Mémoire présenté le : 23/05/2023

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Adeline Jelsch--Bisel

Titre : Evaluation du risque Cyber des établissements de santé

Confidentialité : ☐ NON ☒ (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires

B. Pierson

C. Meunier

F. Picard

F.X. Négri

Membres présents du jury de l'ISFA

E. Masiello

D. Clot

Signature

Entreprise :
Nom : Relyens

Signature : Laurence Rameaux

Directeur de mémoire en entreprise :
Nom : Alban Garnier

Signature :

Invité :
Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : risque Cyber, tarification, données publiques, secteur de la santé, violations de données personnelles, régression logistique, déséquilibre des classes, rééchantillonnage, ajustement de loi de probabilité.

L'assurance Cyber est un marché en croissance, qui cherche à répondre au nombre grandissant d'incidents et d'attaques informatiques, en particulier dans certains secteurs comme celui de la santé. La modélisation du risque, évolutif dans le temps et aux conséquences difficiles à appréhender, est un défi pour les assureurs.

Afin de proposer une solution à ses sociétaires du secteur de la santé, Relyens commercialise une assurance Cyber, dont le tarif repose sur un modèle actuariel, ainsi que des questions mesurant la sécurité informatique.

Dans l'objectif de gagner en expertise Cyber, et de challenger la tarification actuelle, un modèle alternatif est construit dans ce mémoire. Le risque Cyber étant récent, Relyens ne dispose que d'un historique de sinistralité restreint. Des données publiques américaines seront donc utilisées pour la modélisation, avec application de différents retraitements pour s'adapter au contexte du contrat d'assurance de Relyens et au marché français.

Dans ce mémoire, une régression logistique est effectuée pour modéliser la probabilité de survenance, avec des techniques de rééchantillonnage pour remédier au déséquilibre des classes se traduisant par une faible proportion d'incidents Cyber dans les données. Une loi est ajustée au logarithme du nombre d'individus affectés. Enfin, la probabilité de déclenchement des garanties du contrat et leur coût sont déterminés par avis d'expert.

Abstract

Keywords : Cyber risk, pricing, public data, health sector, data breach, logistic regression, class imbalance, sampling, probability distribution fitting.

Cyber insurance is a growing market, which aims at responding to the increasing number of Cyber incidents and attacks, especially in some sectors such as healthcare. Modeling Cyber risk, which evolves over time and whose consequences are difficult to apprehend, is a challenge for insurers.

In order to offer a solution to its clients in the health sector, Relyens markets a Cyber insurance, whose price is based on an actuarial model, as well as questions measuring IT security.

With a view to gaining Cyber expertise, and to challenging the current pricing, an alternative model is built in this study. Cyber risk being a recent phenomenon, Relyens has only a limited loss history. American public data will therefore be used for the modeling, with the application of some adjustments to adapt to the context of Relyens' insurance contract and to the french market.

In this paper, a logistic regression is performed to model the probability of occurrence, with resampling techniques to address the imbalance of classes resulting in a low proportion of Cyber incidents in the data. A distribution is fitted to the logarithm of the number of affected individuals. Finally, the probability of triggering the contract's guarantees and their cost are determined by expert opinion.

Remerciements

Je voudrais remercier le groupe Relyens, le directeur général Dominique Godet, le directeur du département Assurance Sabri Boudrama et le manager de l'équipe Actuariat Non-Vie Jérôme Schaeffer pour m'avoir offert l'opportunité de réaliser ce mémoire dans le cadre de mon altenance.

Je souhaite adresser mes profonds remerciements à mon tuteur d'entreprise, Alban Garnier, pour ses conseils avisés, sa disponibilité et ses explications qui ont rendu ce mémoire possible.

Je tiens également à remercier les collaborateurs de Relyens qui m'ont beaucoup aidé en apportant leur expertise en risque informatique, en assurance Cyber et en Data Science, en particulier Christophe, Lionel, Arnaud, Pasquale, Thorsten, Laura, Léa et Guilhem.

Je remercie également ma tutrice académique Anne Eyraud-Loisel pour ses conseils et son accompagnement, ainsi que Olivier Lopez qui s'est rendu disponible pour me donner ses conseils au début de la rédaction du mémoire.

Enfin, mes pensées vont à ma famille et à mes amis qui m'ont soutenue dans ce projet.

Sommaire

Introduction

Dans le contexte actuel marqué par une forte digitalisation et une interconnexion numérique croissante, les administrations et entreprises sont de plus en plus victimes d'incidents Cyber. Les cyberattaques qui ont touché l'hôpital de Dax [1], ou bien celui de Villefranche-sur-Saône [2], sont deux exemples récents survenus en 2021 qui montrent que les attaques n'épargnent pas le secteur de la santé. En effet, les structures de soins médicaux sont particulièrement vulnérables aux incidents informatiques car elles disposent de parcs informatiques importants, et également de nombreux objets connectés et dispositifs médicaux. De plus, face à l'urgence d'assurer la continuité des soins, un établissement de santé peut être davantage susceptible de répondre à la demande des pirates qu'une entreprise classique. Par exemple, il pourrait payer plus rapidement la rançon demandée par un rançongiciel pour retrouver au plus vite l'accès aux données patients. Les établissements de santé constituent donc une cible privilégiée pour les pirates informatiques. D'après l'ANSSI (Agence Nationale de la Sécurité des Systèmes d'Information), 11% des victimes de rançongiciels en 2020 sont des établissements du secteur de la santé. [3]

Face à ce risque croissant qui devient une préoccupation pour les établissements de soins, la demande en assurance, en service de protection et d'assistance face au risque Cyber est forte. Le groupe Relyens est un acteur de référence en assurance à destination des établissements et professionnels de santé. L'entreprise propose depuis 2017 une solution d'assurance et de prévention aux établissements de santé et médico-sociaux qui leur permet de se prémunir des conséquences d'un incident Cyber.

La tarification de cette assurance se base aujourd'hui sur un modèle qui est calibré sur des données concernant plusieurs secteurs d'activité et régions du monde. La prime est différenciée en fonction du pays et de la catégorie d'entreprise considérés. Cependant, il est possible d'imaginer que les valeurs des primes sont influencées par des secteurs ou régions du monde sur-représentées dans les données ayant servi à la construction du modèle. La prime issue du modèle actuel dépend également fortement d'un questionnaire de risque basé sur la norme ISO 27001. Relyens réalise une évaluation du risque additionnelle afin d'éventuellement ajuster la prime issue de ce modèle. Ainsi, les actuaires travaillent en collaboration avec des experts en sécurité informatique afin de développer des solutions d'analyse du risque spécifique au secteur de la santé et de trouver des moyens de le quantifier. Aujourd'hui, un questionnaire supplémentaire est proposé aux sociétaires, construit pour analyser la sécurité informatique d'une structure de soins. Ce mémoire s'inscrit donc dans la démarche de Relyens de mieux s'approprier le risque, en gagnant en expertise sur son évaluation.

Dans ce contexte, l'objectif de ce mémoire est de construire par une approche actuarielle un modèle alternatif d'évaluation du risque Cyber, adapté au secteur de la santé et au marché français, afin de challenger le modèle actuel. De plus, la prime issue du modèle actuel dépend fortement du questionnaire

1. Source : https://www.ticsante.com/story?ID=6141, article publié le 08 avril 2022
2. Source : https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/ce-que-l-on-sait-de-la-cyberattaque-contre-l-hopital-de-villefranche-sur-saone_4299065.html, article publié le 16 février 2021
3. Source : https://www.ssi.gouv.fr/actualite/cybersecurite-faire-face-a-la-menace-la-strategie-francaise/

de risque informatique qui peut être assez volumineux. Le mémoire vise donc également la création d'un tarificateur qui peut apporter un gain de temps de souscription.

L'historique de sinistralité du groupe est trop restreint pour construire un modèle actuariel, et il n'existe pas de base de données à l'échelle nationale ou européenne concernant ce risque. Ainsi, des données publiques américaines seront utilisées. Les données de violations de données personnelles de santé publiées par *HHS Office for Civil Rights (U.S. Department of Health and Human Services Office for Civil Rights)*[4] regroupent les déclarations que sont tenus de faire les établissements en cas d'incident mettant en danger la confidentialité des données personnelles de santé, au-dessus d'un certain seuil d'individus concernés. Elles serviront de données sinistres. La fréquence de déclaration des hôpitaux américains sera recherchée à partir d'une base publiée par *CMS (Center for Medicare and Medicaid)* qui jouera le rôle de la base contrats. Elle répertorie des informations structurelles et financières concernant plus de 6000 hôpitaux américains[5]. Les hôpitaux de cette base seront donc associés à leurs éventuelles déclarations d'incidents.

Plusieurs catégories d'incidents Cyber sont identifiables et sont modélisées de manière distincte. La probabilité de déclaration des hôpitaux américains est modélisée à l'aide de régressions logistiques et techniques de rééchantillonnage. Elle sera assimilée à une fréquence puis ajustée selon le nombre de déclarations sous le seuil d'individus concernés, et le nombre de déclarations faites en France, afin d'approximer la fréquence réelle d'incidents pour les établissements de santé français. Le coût n'est pas disponible dans les données étudiées, il est donc déterminé par avis d'expert en se basant sur le type d'incident, le nombre d'individus affectés, et des caractéristiques de l'établissement touché. Le logarithme du nombre d'individus affectés quant à lui fait l'objet d'une modélisation par ajustement d'une loi normale et de Gumbel. Une matrice de coûts sera finalement construite, regroupant un coût par garantie et type d'incident. Cela permet ainsi la construction d'un modèle de tarification inspiré du traditionnel modèle fréquence $\times$ coût.

Dans un premier temps, le contexte français, européen et américain du risque Cyber est présenté d'un point de vue assurantiel et réglementaire, afin de mettre en évidence les principaux enjeux spécifiques au secteur de la santé. Dans un second temps, une première analyse des données des incidents Cyber et des hôpitaux américains est réalisée. La troisième partie porte sur la modélisation du risque en se basant sur les données américaines. Enfin, la dernière partie décrit la modélisation effectuée et les ajustements réalisés afin d'aboutir à un modèle adapté au marché français et aux sociétaires de Relyens.

4. Base disponible sur le site : https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf
5. Base disponible sur le site : https://data.cms.gov/provider-compliance/cost-report/hospital-provider-cost-report

Chapitre 1

Présentation du risque Cyber et de ses enjeux pour le secteur de la santé

1.1 Présentation du risque Cyber

1.1.1 Définition du risque Cyber

Le risque Cyber d'une organisation désigne de manière générale la probabilité qu'un évènement survienne et porte atteinte à son système d'information et/ou à ses données numériques. L'incident peut être le résultat d'un acte malveillant ou accidentel [6]. Selon le *CRO (Chief Risk Officers) Forum*, le risque Cyber correspond à « tous les risques découlant de l'utilisation des données électroniques et de leur transmission, y compris les outils technologiques tels qu'Internet et les réseaux de télécommunications. Cela englobe également les dommages physiques pouvant être causés par des incidents de cybersécurité, la fraude commise par une mauvaise utilisation des données, toute responsabilité découlant du stockage des données et la disponibilité, l'intégrité et la confidentialité des informations électroniques - qu'elles soient liées à des individus, des groupes ou des gouvernements » [7]. L'ANSSI (Agence Nationale de la Sécurité des Systèmes d'Information) définit dans son glossaire un incident de Cyber-sécurité comme un « évènement qui porte atteinte à la disponibilité, la confidentialité ou l'intégrité d'un bien ».

Ainsi, le risque regroupe une grande variété d'incidents et de conséquences possibles. Cependant, cette variété est réduite par le contrat d'assurance qui défini les évènements couverts et d'éventuelles exclusions.

1.1.2 Principales catégories d'incidents et leurs conséquences

Généralement, deux grandes catégories d'incidents sont distinguées : les actes malveillants qui peuvent être d'origine externe ou interne à l'organisation, et les causes accidentelles. Concernant l'incident intentionnel, les pirates informatiques sont des individus extérieurs à l'organisation qui parviennent à s'introduire dans le système informatique. La malveillance peut également être interne, par exemple un acte commis par un employé. Lorsque les incidents sont accidentels, les causes peuvent être variées comme un dysfonctionnement informatique, ou encore une erreur humaine.

6. Christophe Delcamp, Présentation *Les risques numériques et la cyber assurance*, Support de formation « La gestion des risques cyber de la prévention à la couverture assurantielle » délivrée par l'Institut du Risk Management

7. *CRO Forum*, *Cyber resilience The cyber risk challenge and the role of insurance*, Décembre 2014, définition traduite

Le risque Cyber ayant pris de l'ampleur ces dernières années, l'ANSSI est l'autorité nationale dont le rôle est d'accompagner le développement du numérique et d'assurer sa sécurité. Elle surveille particulièrement le risque malveillant externe et décrit quatre menaces pour les entreprises et administrations : la déstabilisation, le sabotage, l'espionnage et la cybercriminalité. Chaque menace est associée à un ou plusieurs types de cyberattaques :

— La déstabilisation peut être provoquée par une attaque par déni de service qui vise à rendre un service indisponible en le saturant. Elle peut aussi être causée par la modification d'un site internet, le vol et la divulgation de données.

— Le sabotage survient après une attaque rendant inopérant le système d'information de l'entité.

— L'espionnage peut survenir après une attaque par point d'eau. Il s'agit d'utiliser un site Internet comme point d'entrée pour infecter des ordinateurs. Une autre cause possible est l'hameçonnage ciblé qui désigne l'envoi de mails contenant des virus.

— La cybercriminalité a un but principalement lucratif. Très médiatisé, le rançongiciel correspond à un virus qui perturbe le fonctionnement du système d'information ou empêche l'accès à des données. Un paiement est demandé afin de rétablir l'accès aux données ou le fonctionnement du système. Une autre forme de cybercriminalité est l'attaque par hameçonnage qui consiste en l'envoi de mails demandant d'indiquer des données personnelles dans le but de dérober de l'argent, ou des identifiants permettant l'accès à certains systèmes d'information.

Il est donc possible d'imaginer de multiples conséquences financières qu'un assureur pourrait prendre en charge dans une police d'assurance dédiée. Les dommages matériels, immatériels, la perturbation de l'activité, les moyens mis en oeuvre pour remédier à l'incident et une éventuelle extorsion entraînent des pertes financières directement liées à l'incident. En cas de préjudice causé à des tiers, la responsabilité d'une personne morale peut être mise en cause et sa réputation peut être dégradée. Ces conséquences résultent en pertes financières indirectes. Face à la variété des causes et conséquences liées aux outils informatiques et aux données numériques, les assureurs délimitent précisément le risque au sein du contrat d'assurance en définissant les garanties couvertes et en mentionnant d'éventuelles exclusions.

L'EIOPA (Autorité européenne des assurances et des pensions professionnelles) décrit les garanties les plus fréquentes proposées par les assureurs et réassureurs dans son enquête *Cyber Risk for Insurers - Challenges and Opportunities* publiée en 2019 [8]. Dans l'ordre de la plus fréquente à la moins fréquente, il s'agit de :

— La restauration des données
— Le vol Cyber ou perte financière entraînée par le transfert électronique frauduleux de fonds
— Les dommages causés aux tiers
— Les coûts de nettoyage du système
— Les incidents de données électroniques liés à un dommage accidentel du système informatique
— La Cyber extorsion qui peut prendre en compte le coût de la rançon ou l'intervention d'un expert
— Les frais administratifs d'investigation et amendes administratives
— La perte d'exploitation, dommages matériels ou immatériels associés à l'interruption du réseau

8. Cette nomenclature de garanties est aussi présentée dans le rapport de l'EIOPA *Understanding Cyber Insurance* de 2017 de manière plus détaillée

— Les premières réponses à l'incident qui peuvent être la gestion de crise et l'intervention d'experts, une aide juridique, ou encore les coûts d'investigation
— Les dépenses supplémentaires

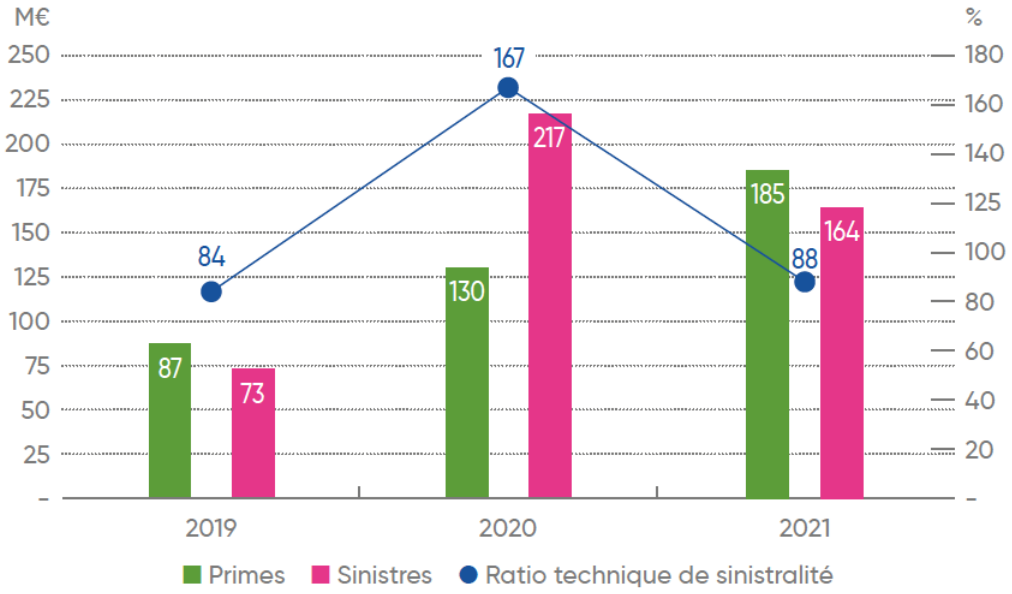
55% des participants ont mentionné d'autres garanties, ce qui illustre la variété des produits proposés en Europe.

Les assureurs et réassureurs doivent également être attentifs au risque de couverture silencieuse. Il désigne le risque qu'un évènement Cyber survienne et engendre des dommages déclenchant des garanties d'assurance dont le Cyber n'était pas la cible, par exemple une garantie dommages aux biens ou responsabilité civile.

1.1.3 Marché de l'assurance Cyber en France et aux Etats-Unis

Deux études LUCY (LUmière sur la CYberassurance) de l'AMRAE (Association pour le Management des Risques et des Assurances de l'Entreprise) ont été publiées en 2021 et en 2022. Elles décrivent le marché de l'assurance Cyber en France de 2019 à 2021.

Le graphique suivant issu de l'étude publiée en 2022 présente les montants de primes et de sinistres en millions d'euros, ainsi que les ratios Sinistres / Primes, entre 2019 et 2021.



Source : Étude LUCY menée par l'AMRAE en 2021.

FIGURE 1.1 – Résultats techniques de l'assurance Cyber

Une forte dégradation du ratio Sinistres / Primes a eu lieu entre 2019 et 2020. En effet, malgré une hausse des primes, les indemnisations ont augmenté encore plus fortement, en raison de 4 sinistres majeurs ayant touché des entreprises de plus de 1,5 milliard d'euros de chiffre d'affaires. Ils représentent uniquement 1% des sinistres Cyber en nombre en France en 2020, mais 78% des indemnisations versées.

Le volume des primes a continué d'augmenter en 2021, malgré une baisse des capacités souscrites. En effet, l'étude publiée en 2022 indique que le marché en 2021 est rentable grâce au ratio Sinistres /

Primes des grandes entreprises qui a fortement baissé pour atteindre 58%. Cette amélioration peut en partie s'expliquer par le durcissement des conditions de souscription et de renouvellement, et une forte hausse des franchises. Il est en revanche dégradé pour les petites et moyennes entreprises.

Les études mettent également en avant que le marché de la Cyber-assurance est particulièrement développé pour les grandes entreprises qui ont un taux de couverture de 87% et représentent 82% du volume des primes collectées en 2020. Ce taux de couverture a cependant baissé de 4,4% en 2021, probablement en raison du durcissement des conditions de souscription. Au contraire, malgré une hausse en 2020, le taux de couverture des petites et moyennes entreprises reste faible.

L'étude de 2022 révèle également que les sinistres sont principalement d'origine malveillante, les accidents ne constituant que 5,5% des sinistres et 4% du volume des indemnisations.

Concernant l'assurance Cyber aux Etats-Unis, le rapport *Cybersecurity Insurance Market 2020* de *NAIC (National Association of Insurance Commissioners)* publié en 2021 révèle quelques tendances similaires à la France en 2020. Il indique que le volume des primes souscrites au titre de garanties Cyber augmente depuis 2016 pour atteindre 2,75 milliards de dollars en 2020. Comme en France, les ratios Sinistres / Primes des vingt plus grands assureurs du risque Cyber se sont dégradés entre 2019 et 2020 pour atteindre près de 67% en moyenne, alors que la moyenne était d'environ 45% en 2019.

En synthèse, le marché américain est plus volumineux et plus ancré que celui de la France, mais tous deux ont connu une dégradation de la sinistralité en 2020. Les entreprises semblent toutes exposées au risque Cyber. En France, les plus grandes d'entre elles ont bien conscience du risque et sont très largement couvertes. Néanmoins, du fait de son activité, le monde de la santé est particulièrement exposé, et le risque peut parfois se traduire de manière différente par rapport à d'autres secteurs.

1.1.4 Spécificités du risque pour les établissements de santé

Comme toutes les entreprises, les établissements de soins connaissent une digitalisation et une externalisation de leurs systèmes et données à des tiers. De plus, certains établissements forment des regroupements comme les GHT (Groupements Hospitaliers de Territoire) ce qui se traduit par une forte interconnexion entre les hôpitaux, par exemple avec la mutualisation de fonctions supports comme le système d'information [9].

La spécificité des établissements de santé réside dans leur activité même. En effet, les activités de soins relèvent d'une réglementation particulière, afin d'assurer l'amélioration de la santé des patients, et la continuité des soins. Face à ces obligations découle naturellement une préoccupation tournée vers la sécurité des patients et la disponibilité permanente de l'offre de soins. Le risque informatique peut donc être secondaire dans certaines situations. Par exemple, il sera choisi d'immédiatement débrancher un appareil infecté par un virus informatique pour protéger le patient dont la santé serait liée au bon fonctionnement de cet outil, sans prendre en considération les dommages sur le plan purement Cyber comme la perte de données ou l'endommagement de l'appareil.

Il faut également considérer la diversité du parc informatique dans ces organisations. Il contient à la fois des technologies de l'information appelées *IT (Information Technology)*, de nombreuses technologies opérationnelles *OT (Operational Technology)* et des dispositifs médicaux. Les *OT* désignent la technologie opérationnelle qui est très présente dans le secteur de la santé et est essentielle pour son fonctionnement. Par exemple, la gestion de la qualité de l'air en bloc opératoire dépend de technologies opérationnelles. Face à ce vaste parc de technologies, le nombre d'employés est relativement faible et il

9. Source : https://solidarites-sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/groupements-hospitaliers-de-territoire/

peut être difficile de maintenir un certain niveau de sécurité informatique en cas de manque de moyens, la priorité étant avant tout la sécurité du patient. Les dispositifs médicaux désignent des outils à action mécanique afin d'effectuer un diagnostic ou à but thérapeutique. Il s'agit d'équipements destinés à concourir à l'exercice des activités de médecine via des technologies ayant été approuvées comme sûres par les autorités. De nombreux dispositifs permettent d'alimenter ou de récupérer des informations directement auprès du DPI (dossier patient informatisé) afin d'optimiser la chaîne de soins et limiter les erreurs de transmission d'informations ; à ce titre ils sont considérés comme exposés car connectés au réseau. Une particularité réside dans le fait que les établissements ne sont pas autorisés à modifier les dispositifs médicaux, y compris pour y installer un antivirus ou un autre système de protection, car la gestion du risque sur ces équipements incombe principalement aux fabricants. Cependant, ils ne font pas encore aujourd'hui l'objet d'attaques informatiques en France grâce à leur technologie complexe et à la sécurité exigée par la réglementation.

Les établissements de santé font néanmoins face à de nombreux incidents Cyber touchant les technologies de l'information. Les plus médiatisés sont les rançongiciels qui perturbent le fonctionnement du système informatique de l'établissement ou rendent inaccessibles les données et vendent des clés de déchiffrement. La principale motivation des cyberattaques est financière. Les établissements de santé sont particulièrement ciblés par les attaques car ils pourraient être plus susceptibles de payer une rançon devant l'urgence d'assurer la continuité des soins. Cependant, payer la rançon n'est pas recommandé car rien ne garantit l'intégrité de la donnée récupérée, et cette intégrité est cruciale pour l'activité de soin. Il est craint qu'une attaque puisse perturber le déroulé des soins et atteindre à la sécurité des patients. C'est aujourd'hui heureusement peu le cas, les attaques ciblent principalement des services administratifs dans une motivation purement financière, sans toucher directement à la sécurité ou à la vie des patients car cela mobiliserait fortement les forces de l'ordre ce qui n'est pas l'intérêt du pirate.

Également, les données qu'ils possèdent sont particulièrement sensibles, notamment lorsque certains de leurs patients sont des célébrités. Cet effet est néanmoins à relativiser. D'abord, la valeur marchande d'une donnée personnelle est plus faible en France qu'aux Etats-Unis, et les systèmes juridiques sont différents. De plus, une grande partie des cyberattaques sont automatisées, des robots recherchent des failles de sécurité pouvant permettre l'intrusion d'un pirate informatique. Tous les établissements connectés à un réseau sont donc exposés.

Les graphiques suivants sont issus du rapport public 2021 de l'Observatoire des signalements d'incidents de sécurité des systèmes d'information pour le secteur santé. Ils résument le nombre d'incidents déclarés au CERT Santé (*Computer Emergency Response Team*) par les établissements de santé et médico-sociaux en 2020 et 2021, par type d'origine.

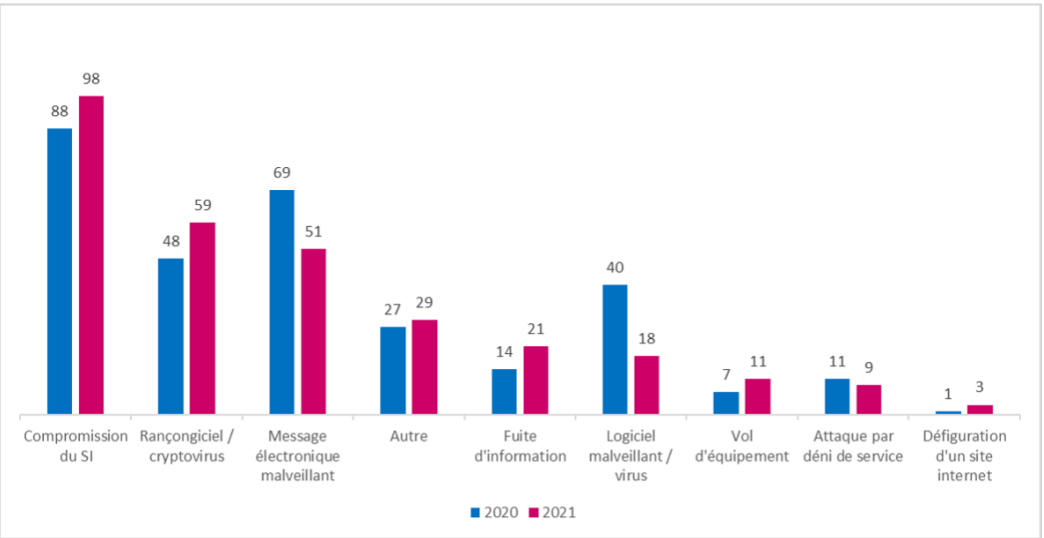


FIGURE 1.2 – Nombre d'incidents par type d'origine malveillante

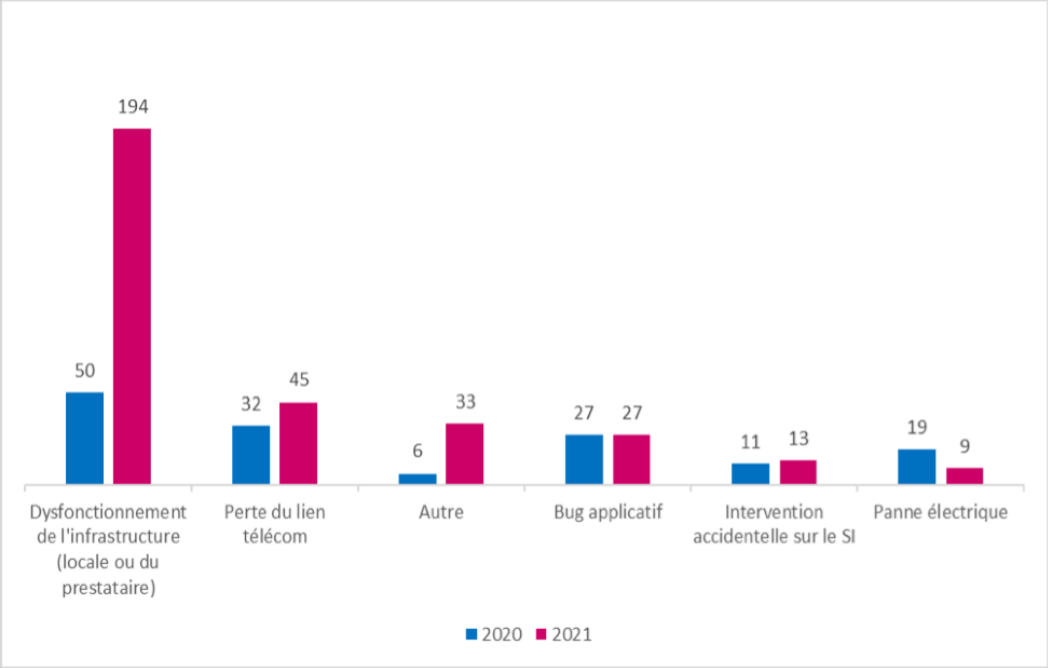


FIGURE 1.3 – Nombre d'incidents par type d'origine non malveillante

La source malveillante la plus fréquente est la compromission du système d'information, qui peut être liée à la récupération d'un compte pour accéder aux systèmes ou bien l'exploitation de vulnérabilités. Le rançongiciel est la seconde attaque la plus fréquente. Concernant les actes non malveillants, la principale cause est un dysfonctionnement de l'infrastructure, le rapport précise qu'il s'agit principalement d'incidents connus par des hébergeurs ou prestataires de solutions *SaaS* (*Software-as-a-Service*) qui désignent des logiciels hébergés sur des serveurs externes.

Enfin, une autre spécificité est le statut d'OSE (Opérateur de Service Essentiel) voire d'OIV (Opérateur d'Importance Vitale) de certains établissements. Les OIV désignent les opérateurs qui utilisent des installations indispensables pour le fonctionnement et la survie du pays, comme le secteur de l'énergie ou de l'alimentation. La liste de ces opérateurs est confidentielle. Une catégorie plus

large désigne les OSE et correspond à un opérateur qui fournit un service dit « essentiel », dont l'interruption aurait des conséquences majeures sur le fonctionnement de l'économie et de la société. Ces opérateurs ont des contraintes réglementaires plus fortes que les autres entreprises. Ce statut, ou plus généralement le secteur d'activité de la santé, peut entraîner une intervention étatique d'ampleur en cas de sinistre. L'ANSSI peut intervenir et formuler des recommandations pour l'amélioration de la sécurité informatique. Cependant, le principe assurantiel correspond à l'indemnisation et à la restauration pour retrouver l'état initial et ne peut donc financer une amélioration.

Ces spécificités peuvent rendre l'analyse du risque et la gestion des sinistres particulièrement longues et complexes, ce qui doit être pris en considération par l'assureur.

1.1.5 Exemples d'incidents Cyber concernant des établissements de santé

Les attaques subies par des hôpitaux ou cliniques en Europe sont parfois médiatisées. Cette section présente quelques incidents et leurs conséquences.

Le CHU (Centre Hospitalier Universitaire) de Montpellier a connu une attaque par hameçonnage en mai 2019 [10]. Un employé a ouvert un mail qui a ensuite contaminé le réseau et touché 649 postes sur 6000. Il n'y a pas eu d'impact sur le secret médical, ni de perturbation majeure des soins car les soignants ont utilisé des ordinateurs portables pendant la désinfection des ordinateurs fixes.

Les attaques peuvent parfois toucher tout le réseau, immobilisant ainsi le système d'information et mettant en difficulté l'établissement. C'est le cas de l'hôpital de Dax qui a connu une cyberattaque majeure. Infecté par le rançongiciel *Ryuk*, la totalité du système d'information a été mis hors service en février 2021. Le jour de l'attaque, tous les accès informatiques ont été coupés et l'hôpital a fonctionné en "mode dégradé" avec l'utilisation de papier. Des ordinateurs non interconnectés ont été mis en place pour accéder aux données sauvegardées, puis après quelques jours, des ordinateurs portables ont été fournis. Le coût total de l'attaque s'est élevé à 2,3 millions d'euros. 174 000 € ont été payés au titre de la reconstruction du réseau, 546 000 € au titre des prestations de cybersécurité et réinstallations, 9 000 € au titre de la sous-traitance biologie, 1,48 million d'euros au titre des coûts du personnel, et enfin 143 000 € au titre des pertes de recettes commerciales. Les coûts ont été pris en charge par l'ARS (Agence Régionale de Santé) Nouvelle-Aquitaine.

L'hôpital de Villefranche-sur-Saône a également été touché par le rançongiciel *Ryuk* bloquant l'accès aux données d'environ 3000 ordinateurs. Les nouveaux patients nécessitant des soins urgents ont été réorientés vers d'autres hôpitaux et cliniques.

Les incidents Cyber peuvent également avoir une cause interne. L'Hôpital Haga de la Haye aux Pays-Bas [11] s'est vu infliger une amende de 460 000 € pour violation des données car certains de ses employés ont consulté le dossier médical d'une célébrité qui a ensuite porté plainte.

A travers ces exemples, le risque Cyber constitue bien une réalité pour le monde de la santé, avec des enjeux multiples mobilisant différentes réglementations et acteurs associés.

10. https://france3-regions.francetvinfo.fr/occitanie/herault/montpellier/chu-montpellier-victime-attaque-informatique-plus-600-ordinateurs-infectes-1670717.html, publié le 17/05/2019
11. Source : https://www.ticsante.com/story?ID=4715 *Informations sur l'incident touchant l'hôpital Haga de la Haye*, publié le 29/07/2019

1.2 Acteurs et contexte réglementaire associés au risque Cyber

1.2.1 Acteurs en France

A l'échelle nationale, deux principaux acteurs sont responsables de la sécurité informatique en France. L'ANSSI est rattachée au secrétaire général de la défense et de la sécurité nationale et apporte son expertise à diverses organisations. Elle dirige également les décisions gouvernementales relatives à la sécurité informatique du pays. La CNIL (Commission Nationale de l'Informatique et des Libertés) veille quant à elle à la sécurité et à la confidentialité des données personnelles.

De plus, un site Internet est mis en place par le gouvernement à destination des particuliers, des entreprises et des collectivités publiques pour sensibiliser au risque Cyber, conseiller sur sa prévention, et porter assistance aux victimes de cyber-attaques. Il livre également des statistiques nationales sur ce risque [12].

Au niveau européen, l'Agence de l'Union européenne pour la cybersécurité (ENISA) participe à la politique européenne de cybersécurité et coopère avec les membres.

Pour le secteur de la santé, L'ANS (Agence du Numérique en Santé) a pour rôle de contribuer au développement du numérique en santé et d'assurer sa sécurité. Sa cellule CERT Santé enregistre et analyse les signalements d'incidents de sécurité et apporte son soutien en cas d'incident.

1.2.2 Contexte réglementaire Cyber en France et en Europe

La réglementation est de plus en plus contraignante dans le but d'imposer un niveau minimum de sécurité informatique aux entreprises et de rendre la notification des incidents obligatoire. Les lois peuvent se regrouper en deux catégories : celles visant à protéger les données personnelles, et celles portant sur la protection du système d'information.

D'abord, les données à caractère personnel ont fait l'objet de plusieurs réglementations. La loi Informatique et Libertés du 6 janvier 1978 indique le droit à l'accès, la modification et la suppression des données personnelles par les individus concernés, mais également le type de données qui peuvent être collectées, l'obligation d'assurer la sécurité des données, et l'obligation d'informer les individus concernés et la CNIL de la collecte des données [13]. Puis, le RGPD (Règlement Général sur la Protection des Données) adopté le 27 avril 2016 et entré en vigueur le 25 mai 2018 a pour objectif d'uniformiser les règles concernant les données personnelles au sein de l'Union Européenne. Il définit des règles de protection et de traitement des données personnelles pour les organisations implantées en Union Européenne, ou dont les services sont à destination de l'Union Européenne.

Concernant la sécurité des systèmes d'information, de nombreuses lois et stratégies ont été mises en place depuis plusieurs années. En 2008, le Livre Blanc sur la défense et sécurité nationale évoque la nécessité de prendre en compte la prévention des attaques informatiques dans la sécurité nationale. Celui de 2013 est concrétisé par la loi de programmation militaire la même année. Il définit les opérateurs vitaux pour la nation (OIV) et introduit des obligations de mesures de sécurité et de déclaration des incidents pour ces entités. Au niveau européen, la directive NIS (Network and Information System Security) du 6 juillet 2016 est une directive européenne qui définit les opérateurs de services essentiels (OSE), élargissant le statut d'OIV. Elle vise à assurer un niveau minimum de sécurité chez les OSE et les FSN (Fournisseurs de Services Numériques). En particulier, les 136 GHT (groupements hospitaliers

12. Lien du site : https://www.cybermalveillance.gouv.fr/
13. Source : https://donnees-rgpd.fr/loi-informatique-libertes/

de territoire) sont considérés comme des OSE [14]. Enfin, la loi relative à la programmation militaire de 2019-2025 vise le renforcement des capacités de détection des attaques informatiques.

La volonté de mieux prévenir ce risque au niveau national se traduit également par des investissements des organismes publics. Le plan France Relance, déployé en 2020 dans le but de favoriser le développement de certains secteurs, comprend un volet cybersécurité de 136 millions d'euros. Un plan à 1 milliard d'euros est aussi déployé en 2021 pour lutter contre la cybercriminalité, par exemple au travers d'investissements dans la recherche en cybersécurité et le renforcement de l'ANSSI. [15].

Ce renforcement se constate aussi au niveau européen. Le règlement européen *Cybersecurity Act* adopté en 2019 met en place un mandat permanent pour l'ENISA, l'Agence européenne pour la cybersécurité et donne une définition d'un cadre européen de certification de cybersécurité.

Les réglementations indiquent également des sanctions possibles en cas de non respect de la sécurité informatique ou de la confidentialité. Le RGPD prévoit que le montant des sanctions pécuniaires peut s'élever à 20 millions d'euros ou 4 % du chiffre d'affaires annuel mondial, et peut être publié. Les sanctions pénales sont de 5 ans d'emprisonnement et 300 000 € d'amende. Le 7 décembre 2020, la CNIL a sanctionné deux médecins par des amendes de 3 000 € et 6 000 € pour ne pas avoir assuré une protection minimum des données de santé de leurs patients et ne pas avoir déclaré un incident. Le 15 avril 2022, la société DEDALUS BIOLOGIE s'est vue infliger une amende de 1,5 million d'euros à cause de défauts de sécurité ayant engendré la fuite de données médicales d'environ 500 000 personnes [16].

1.2.3 Réglementation en France pour le secteur de la santé

Les données de santé font l'objet d'une réglementation particulière. D'après l'article L.1111-8 du Code de la Santé publique, « toute personne physique ou morale qui héberge des données de santé à caractère personnel recueillies à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi médico-social pour le compte de personnes physiques ou morales à l'origine de la production ou du recueil de ces données ou pour le compte du patient lui-même », doit être agréée ou certifiée à cet effet. Mis à part les établissements de santé gérant leur propre système d'information, les organismes qui conservent ou exploitent des données de santé doivent être certifiés Hébergeurs de Données de Santé. La notion de confidentialité des données des patients est également présente dans le Code de la Santé au travers du secret médical. Par exemple, un médecin peut consulter uniquement les données de ses patients. En cas de besoin, il peut cependant accéder aux dossiers des autres patients ce qui est appelé « bris de glace ». Ces bris de glaces sont enregistrés et il est vérifié que le médecin avait une raison valable pour accéder aux données.

Les dispositifs médicaux font également l'objet d'une réglementation spécifique avec le marquage CE (Conformité Européenne). En cas de panne ou de situation anormale, l'appareil doit pouvoir se mettre en *failsafe*, ce qui se traduit par un signal de l'anomalie et un réglage automatique afin d'éviter tout danger. Selon l'appareil, ce réglage peut être la mise en arrêt, ou rester utilisable lorsque cela est nécessaire pour les soins les plus urgents.

Face à la hausse des cyberattaques touchant les établissements de soins, le plan de renforcement 2021 de la cybersécurité en santé se traduit pour les établissements de santé par des exigences de sécurité, et il est demandé que 5 à 10% du budget informatique soit consacré à la sécurité informatique.

Enfin, il existe une obligation de déclaration des incidents. D'après l'article L1111-8-2 du Code de

14. Source : https://www.ticsante.com/story?ID=5747
15. Source : https://www.lesechos.fr/tech-medias/hightech/cybersecurite-le-plan-a-1-milliard-de-letat-1291369
16. Le montant des sanctions et publications sont disponibles sur le site de la CNIL

la santé publique, les établissements de santé doivent déclarer tout incident significatif de sécurité des systèmes d'informations. Ce signalement se fait auprès de l'ARS (Agence Régionale de Santé). Lorsque des données personnelles sont concernées, le signalement doit également être réalisé à la CNIL, et aux individus concernés lorsque le risque est jugé élevé. Cela fut le cas après la cyberattaque subie par l'hôpital de Corbeil-Essonnes en août 2022, qui a refusé de payer la rançon, et le pirate a par conséquent divulgué des données personnelles. Un courrier a été envoyé aux employés et patients de l'hôpital, avec une incitation à porter plainte [17].

1.2.4 Acteurs et réglementation du secteur de la santé aux Etats-Unis

Les Etats-Unis disposent de plusieurs réglementations à la fois à l'échelle nationale et à celle des Etats. Comme indiqué en introduction, les données utilisées pour la construction du modèle dans ce mémoire sont celles des Etats-Unis. Il s'agit donc d'avoir une vision globale de la réglementation qui s'applique aux Etats-Unis en termes de risque Cyber pour les établissements de santé.

La première loi concernant les données médicales à l'échelle des Etats-Unis date de 1996 et est nommée *Health Insurance Portability and Accountability Act (HIPAA)*. Elle vise à rendre les systèmes de santé plus performants et crée en parallèle des règles de protection des données médicales. Elle est complétée avec d'autres standards de sécurité entre 2000 et 2003 avec la loi *HIPAA Privacy Rule* qui donne également des droits aux individus sur leurs données personnelles et *HIPAA Security Rule*. La loi portant sur les conditions de notifications d'un incident est créée en 2009 avec *The Health Information Technology for Economic and Clinical Health (HITECH)* et est finalisée en 2013 par *HIPAA Omnibus Rule* [18].

Les amendes pour non respect de cette loi varient selon plusieurs critères comme le nombre d'individus affectés, la nature des données concernées, la durée de l'incident, et le niveau de responsabilité de l'entreprise qui peut ne pas avoir eu connaissance d'un manquement ou bien l'avoir volontairement négligé. En 2022, l'amende minimum est de 120 $ minimum et peut aller jusqu'à environ 1,9 million de dollars. [19].

Une réglementation s'applique aussi aux dispositifs médicaux. La *FDA (U.S. Food and Drug Administration)* impose aux fabricants et aux établissements de déclarer tout incident d'un dispositif médical pouvant être à l'origine d'une blessure importante ou du décès d'une personne, ce qui peut être une cause informatique. [20]

17. Source : https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/cyberattaque-a-l-hopital-de-corbeil-essonnes-patients-et-membres-du-personnel-incites-a-porter-plainte_5407198.html
18. Source : https://privacyrights.org/resources/health-insurance-portability-and-accountability-act *(histoire de la réglementation HIPAA)*
19. Source : https://www.hipaajournal.com/what-are-the-penalties-for-hipaa-violations-7096/
20. Source : https://www.fda.gov/medical-devices/postmarket-requirements-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities

1.3 Offre Cyber de Relyens

Les clients de Relyens qui se voient proposer l'assurance Cyber sont principalement des établissements de santé qui appartiennent au périmètre Responsabilité Civile Médicale, mais aussi des acteurs du médico-social et des acteurs territoriaux. Face au risque croissant d'incident Cyber, l'entreprise commercialise une offre d'assurance aux établissements de soins et acteurs territoriaux avec des garanties portant à la fois sur la réparation des dommages et la responsabilité. Elle comprend également une assistance en cas d'incident. L'offre est disponible sur le marché français et s'étend progressivement sur les marchés italien, espagnol et allemand.

Comme d'autres catégories d'assurance, toute souscription à l'assurance Cyber de Relyens doit être précédée d'une évaluation du risque en amont. En effet, les établissements doivent respecter des prérequis, qui correspondent à des niveaux de sécurité et pratiques informatiques minimum pour se protéger du risque. Un établissement jugé inéligible à l'assurance peut être accompagné par Relyens pour augmenter son niveau de sécurité. De plus, afin de compléter son offre, Relyens s'est entouré de différents partenaires. Par exemple, une solution qui peut être proposée en complément de l'assurance ou de manière indépendante a été mise en place avec CyberMDX. Elle a pour but d'évaluer de manière précise la situation informatique de l'établissement afin d'identifier d'éventuelles failles, et ainsi prévenir au mieux le risque.

Actuellement, Relyens dispose d'un historique de sinistralité trop réduit pour construire un modèle de tarification uniquement à partir de ses données. Le tarif actuel se base sur un modèle applicable pour différentes catégories d'entreprises et plusieurs pays du monde, il n'est donc pas spécifique aux établissements de santé et à la France. Il se base sur différentes variables explicatives, ainsi que sur un questionnaire analysant le niveau de sécurité informatique basé sur la norme ISO 27002 (*International Standards Organization*). Ce questionnaire a entre 100 et 200 questions selon la taille de l'établissement.

Les coûts d'un incident Cyber peuvent être très variables selon les catégories d'établissements et les pays pour des raisons réglementaires ou culturelles. C'est pourquoi un questionnaire supplémentaire a été créé par les experts en risque Cyber de Relyens, permettant une prise en compte de variables plus adaptées au monde de la santé. Selon les résultats de ce questionnaire, la prime issue du modèle actuel peut être ajustée par le groupe dans une certaine fourchette. Les travaux de ce mémoire visent à challenger le modèle actuel, grâce à la prise en compte de variables plus ciblées et spécifiques au domaine de la santé. Dans les parties suivantes, un modèle de tarification sera construit pour les établissements de santé en France de manière indépendante aux méthodes de tarification déjà en place au sein de l'entreprise.

En résumé, le risque Cyber dans sa définition élargie peut être lié à une origine malveillante ou non intentionnelle. Les conséquences possibles sont variées, elles peuvent affecter l'assuré même voire des tiers. C'est pourquoi différentes garanties existent généralement dans les contrats d'assurance Cyber, qui définissent précisément les évènements couverts et ceux qui sont exclus. La définition du risque Cyber peut donc varier d'un assureur à un autre.

Il est possible d'observer une multiplication de la médiatisation d'incidents aux conséquences multiples, et une préoccupation croissante des établissements pour s'en prémunir, dont un moyen possible est l'assurance. C'est pourquoi le marché de l'assurance Cyber en France est en croissance depuis 2019, avec un fort taux de couverture des grandes entreprises qui sont les plus conscientes de ce risque.

Les établissements de santé font partie des secteurs les plus touchés et sont donc également conscients du risque. Les incidents les plus déclarés en 2021 au CERT Santé sont la compromission du système informatique, le dysfonctionnement de l'infrastructure et le rançongiciel.

Ainsi, face à une menace croissante, le risque Cyber est très surveillé à l'échelle nationale avec une réglementation de plus en plus rigoureuse en termes de sécurité informatique et de protection des données personnelles. Le secteur de la santé du fait de son activité connaît des réglementations supplémentaires et des spécificités face au risque.

Dans ce contexte, Relyens propose une assurance aux établissements de santé, ainsi que des services de prévention du risque et d'assistance en cas de sinistre. Les tarifs actuels sont issus du modèle qui couple une modélisation actuarielle et un questionnaire de sécurité informatique. Ce tarif est ensuite ajusté par des questions additionnelles créées par les experts en risque informatique de Relyens, plus adaptées au secteur de la santé. L'objectif de ce mémoire est de créer un modèle de tarification alternatif afin de gagner en expertise sur ce risque et de challenger le modèle actuel.

Chapitre 2

Analyse des données disponibles sur les incidents Cyber d'établissements de santé

L'historique de sinistralité interne à Relyens est encore restreint et ne permet pas à lui seul la construction d'un modèle d'évaluation du risque Cyber des établissements de soins. Il était donc nécessaire de trouver d'autres sources de données pour étudier et modéliser le risque. Une recherche des données publiques disponibles sur le risque Cyber a été effectuée ainsi qu'une première analyse afin de déterminer si elles pourraient être utilisées afin de modéliser le risque des établissements de santé.

2.1 Présentation des sources de données publiques d'incidents Cyber

Différentes sources de données publiques concernant le risque Cyber sont disponibles. D'abord, de nombreuses études sont publiées et peuvent être source d'informations importantes. Par exemple, le rapport de l'Observatoire des signalements d'incidents de sécurité des systèmes d'information pour le secteur santé indique des statistiques globales annuelles sur les évènements Cyber subis chaque année par les établissements de santé en France.

Pour des raisons de sécurité et par réticence des entreprises et établissements, en France et en Europe, les évènements Cyber répertoriés par les autorités ne sont pas publiés et seules des statistiques macro sont présentées. La situation est différente aux Etats-Unis où certains incidents Cyber sont publiés sous la forme de bases de données. Ci-dessous sont présentées les bases de données américaines ou internationales trouvées lors des recherches dans le cadre de ce mémoire, et qui peuvent potentiellement faire l'objet d'une étude actuarielle. Le périmètre de chaque base de données est brièvement présenté, et son adéquation avec l'objectif de ce mémoire est discuté.

— La base *Chronology of Data Breaches* de *Privacy Rights Clearinghouse*[21] a souvent été étudiée par des membres de l'Institut des Actuaires pour évaluer le risque Cyber, par exemple dans le mémoire *Etude Actuarielle du Cyber-Risque* de Florian Pons, ou encore le mémoire *Modélisation assurantielle du risque cyber* d'Anaïs Martinez. Elle répertorie les incidents de violations de données pour une grande variété de catégories d'établissements comme les établissements financiers. Les données proviennent de plusieurs sources, par exemple les médias, ce qui permet d'avoir connaissance d'un nombre assez important de violations de données. Cependant, elle n'a pas été mise à jour depuis 2019 ce qui rend plus difficile l'étude du risque Cyber, très évolutif

21. Base disponible sur le site : https://privacyrights.org/data-breaches

qui a pris une forte ampleur ces dernières années. Également, les catégories d'établissements sont peu détaillées et ne permettent pas de distinguer les établissements de soins des autres entreprises du monde de la santé, comme les assurances maladies.

— Une autre source d'informations est la base *Veris Community Database*[22] qui est formée de différentes catégories d'incidents Cyber et dont les données sont anonymisées : les noms des établissements ne sont pas indiqués. Cependant, il n'y a pas d'obligation de déclaration ce qui rend difficile l'étude de la fréquence des incidents à partir de cette base.

— La base de données *MAUDE (Manufacturer and User Facility Device Experience)*[23] publie tous les incidents concernant les dispositifs médicaux aux Etats-Unis. Une faible part d'entre eux correspond à des problèmes de logiciels. Il n'existe pas de variables permettant de les sélectionner, une recherche dans le descriptif serait nécessaire pour les identifier. Les incidents sur les dispositifs médicaux sont aujourd'hui très peu vus et ne constituent pas le risque principal du contrat d'assurance proposé par le groupe Relyens. Cette base n'est donc pas retenue.

— Le portail des violations de données de *HHS OCR* est une base de données qui regroupe des incidents de données personnelles de santé. Elle présente plusieurs avantages pour la construction d'un modèle. En effet, le nombre de données est faible mais suffisant. La base regroupe 3392 incidents d'établissements de santé. Les données sont très récentes car la base est continuellement complétée avec les derniers incidents déclarés. De plus, cette base permet de cibler avec une variable les établissements de soins qui sont la principale cible du produit d'assurance Cyber proposé par Relyens. La déclaration des incidents est obligatoire, il y a donc une quasi exhaustivité du périmètre concerné par l'obligation, ce qui permet une étude de la fréquence.

Ainsi, la base de données de *HHS OCR* répond au mieux à l'objectif de modélisation du risque Cyber des établissements de santé, et est donc retenue pour la construction du modèle. La section suivante présente de manière plus précise le contenu de cette base.

2.2 Présentation de la base d'incidents *HHS Office for Civil Rights breach portal*

Cette section est une étude préliminaire à la modélisation des données de *HHS OCR*. Le périmètre des incidents déclarés est analysé afin de sélectionner pour la modélisation les incidents qui peuvent correspondre à un sinistre selon les termes du produit d'assurance de Relyens, mais également de savoir quelles catégories de sinistres couvertes par l'offre ne peuvent pas être représentées par ces données. Une première analyse exploratoire est réalisée afin de recueillir des informations sur le risque Cyber.

2.2.1 Périmètre des incidents déclarés

La base de données *HHS OCR breach portal* répertorie les incidents de violations de données personnelles de santé qui ont été déclarés à *HHS OCR*. Les données personnelles de santé désignent les informations qui rendent possible l'identification d'un individu et qui sont liées à une activité médicale.

22. Base disponible sur le site : http://veriscommunity.net/vcdb.html
23. Base disponible sur le site : https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/about-manufacturer-and-user-facility-device-experience-maude

Il s'agit par exemple du nom, du numéro de sécurité sociale, d'informations médicales ou de données de facturation.

La déclaration est soumise à la réglementation *HIPAA* 45 CFR §§ 164.400-414 qui définit ce qu'est une violation de donnée et précise les conditions de déclaration. Elle indique également quels établissements sont soumis à cette obligation.

Une violation de données ou incident *Data Breach* est défini selon la réglementation comme une utilisation non autorisée ou une divulgation qui peut porter atteinte à la sécurité ou à la confidentialité d'une donnée personnelle de santé. Cependant, lorsque l'établissement est capable de prouver que la probabilité d'atteinte à la sécurité ou à la confidentialité est très faible, l'incident ne doit pas être déclaré. C'est également le cas lorsqu'un employé a accédé de manière involontaire aux données, si la divulgation a été faite à une personne autorisée, ou bien quand il est fort probable que la personne non autorisée n'a pas pu conserver les données. Il faut également noter que les évènements touchant des données chiffrées ne sont pas considérés comme des incidents et ne doivent donc pas être déclarés.

Cette obligation de déclaration, dont l'autorité de contrôle associée est *HHS OCR*, s'applique aux catégories d'établissements suivants :

— Etablissements de santé (*Healthcare Providers*)
— Régimes de santé et assurances maladies (*Health Plans*)
— Entreprises participant au traitement des données confidentielles de santé (*Health Care Clearinghouses*)
— Prestataires externes de services aux établissements cités précédemment, ayant accès aux données personnelles de santé (*Business Associates*)

La déclaration doit être réalisée dans les 60 jours et elle est publiée lorsque l'incident dépasse le seuil de 500 individus affectés. Après un incident, la déclaration s'accompagne d'autres obligations. Par exemple, l'établissement doit mettre en place pendant au moins 3 mois une ligne téléphonique disponible pour répondre aux questions des individus. Une politique écrite concernant la notification des incidents doit être présente au sein de l'entreprise ainsi qu'une formation des employés à son sujet. Les employés ne la respectant pas doivent être sanctionnés.

2.2.2 Description des variables

Les variables présentes dans la base sont les suivantes :

— *Name of Covered Entity* : Nom de l'établissement
— *State* : Etat fédéré
— *Covered Entity Type* : Type d'entreprise parmi celles décrites ci-dessus
— *Individuals Affected* : Nombre d'individus affectés par l'incident
— *Breach Submission Date* : Date de déclaration de l'incident
— *Type of Breach* : Type d'incident dont les modalités sont les suivantes :
  — *Hacking/IT Incident* : Piratage ou incident informatique
  — *Improper Disposal* : Elimination incorrecte des données
  — *Loss* : Perte des données
  — *Theft* : Vol de données
  — *Unauthorized Access/Disclosure* : Accès non autorisé aux données ou divulgation des données
  — *Other* : Autre
  — *Unknown* : Inconnu

  Ces modalités peuvent se combiner, c'est à dire qu'une ligne peut correspondre à plusieurs

types d'incident.
- *Location of Breached Information* : Localisation des données
  - *Desktop Computer* : Ordinateur de bureau
  - *Electronic Medical Record* : Dossier médical électronique (DME)
  - *Email* : Mail
  - *Laptop* : Ordinateur portable
  - *Network Server* : Serveur réseau
  - *Other Portable Electronic Device* : Autre appareil électronique portable
  - *Paper/Films* : Papier ou CD
  - *Other* : Autre

  Les modalités peuvent ici aussi se combiner.
- *Business Associate Present* : Présence du prestataire externe dans l'incident
- *Web Description* : Description de l'incident

Trois variables peuvent être extraites de la variable Description (non stucturée) par recherche de mots clés [24] :

- Type(s) des données concernées : cliniques, démographiques ou financières.

- Sous-catégorie de l'incident : rançongiciel, hameçonnage, logiciel malveillant (*malware*), erreur non intentionnelle, accès non autorisé, négligence, perte d'un ordinateur de bureau, perte d'un ordinateur portable, perte autre, vol d'un ordinateur de bureau, vol d'un ordinateur portable, vol autre.

- Réponse à l'incident : formation des employés, sanction des employés impliqués, mise en place de nouvelles procédures, changement des mots de passe, mise en place de nouvelles technologies de protection, prestations de protection contre l'ursupation d'identité et de surveillance du crédit aux individus affectés, analyse du risque, a bénéficié de l'assistance d'OCR, mise en place de moyens de protection physique, adoption de technologies de chiffrage, réalisation d'enquêtes, mise en place de nouveaux plans de gestion du risque, modification du contrat avec le partenaire, fin du contrat avec le partenaire, implémentation d'évaluations périodiques.

Les mots clés des variables Type(s) des données concernées et Réponse à l'incident ont été construits en s'inspirant du questionnaire rempli par les établissements pour déclarer un incident [25]. Afin d'aboutir à des modalités quasi exhaustives, les mots et couples de mots les plus fréquents non utilisés en mots clés sont observés dans les descriptions et ajoutés s'ils correspondent à une variable, cela jusqu'à ce que aucun mot fréquent ne semble lié aux variables ci-dessus.

2.2.3 Périmètre des incidents étudiés et retraitements avant analyse exploratoire

Seuls les incidents des établissements de soins sont étudiés [26]. Ceux déclarés en 2022 ne sont pas retenus afin d'étudier des années complètes. *HHS OCR* distingue les incidents encore en cours d'investigation de ceux dont l'investigation est terminée. Les évènements sont étudiés ici indépendamment de cette distinction. En effet, les incidents qui font l'objet d'une publication ont très probablement déjà fait l'objet d'une première vérification. Également, les incidents dont la localisation indiquée est

24. Les mots clés utilisés sont présentés en annexe B, ainsi qu'un nuage des mots les plus fréquents dans la variable Description

25. Formulaire de déclaration d'un incident : https://ocrportal.hhs.gov/ocr/breach/doc/Breach_Portal_Questions_508.pdf

26. Quelques établissements dont la catégorie était mal renseignée sont ajoutés manuellement

le papier uniquement ne correspondent pas à un incident Cyber et sont donc retirés.

Près de 98,6% des incidents ont une unique catégorie d'incident renseignée. Afin de faciliter l'analyse, une seule catégorie est retenue pour chaque incident, avec en priorité le piratage, la divulgation/accès non autorisé, le vol, puis la perte.

2.2.4 Première analyse exploratoire

2757 incidents informatiques ont été déclarés entre 2010 et 2021 par des établissements de soins. Le diagramme en barres 2.1 illustre le nombre d'incidents par année de déclaration et par catégorie. Pour des raisons de lisibilité, les valeurs des effectifs annuels de chaque catégorie sont affichées si elles sont supérieures à 15.

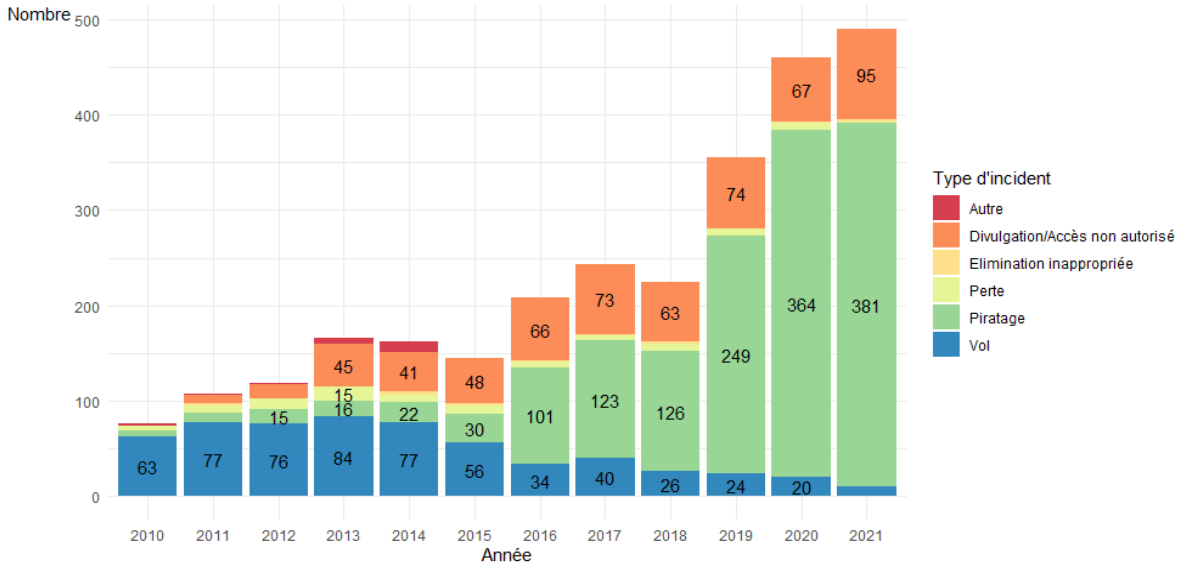


FIGURE 2.1 – Nombre d'incidents par année et par catégorie

Le nombre de piratages est en forte hausse et est depuis plusieurs années la catégorie d'incident la plus fréquente. Au contraire, les vols ont fortement chuté, et les pertes sont des incidents très peu fréquents. Les éliminations inappropriées sont quasi inexistantes au sein des incidents Cyber car elles concernent surtout des données sur papier qui ont été supprimées des données étudiées. Le nombre de divulgations et accès non autorisés est marqué par une hausse à partir de 2013, liée au changement réglementaire. En effet, un établissement devait précédemment déclarer un incident lorsqu'il y avait un risque de dommages pour les individus concernés, alors qu'il s'agit maintenant de déclarer tous les incidents de données sauf s'il est possible de prouver que le risque pour les individus est très faible, selon certaines conditions réglementaires. [27].

Pour chaque catégorie, le graphique 2.2 indique le nombre d'incidents concernant un prestataire et ceux concernant directement l'établissement.

27. Source : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3804103/

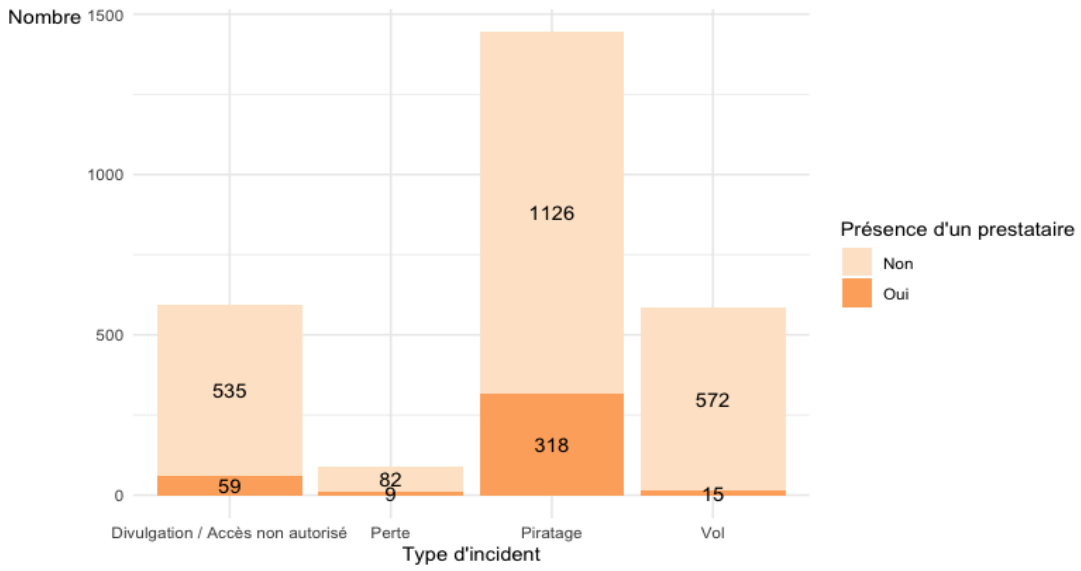


FIGURE 2.2 – Nombre d'incidents avec ou sans implication d'un prestataire

Les incidents subis par les prestataires sont principalement des piratages, et dans une moindre mesure des divulgations ou accès non autorisés.

Le tableau ci-dessous récapitule les plus grands incidents des établissements de santé en nombre d'individus affectés.

Nom	Type d'incident	Nombre d'individus affectés	Année	Description
Laboratory Corporation of America Holdings	Piratage	10 251 784	2019	Piratage du prestataire *Retrieval-Masters Creditors Bureau*
Community Health Systems Professional Services Corporation	Piratage	6 121 158	2014	Les pirates ont utilisé des identifiants compromis pour accéder au système d'information à distance et ont exfiltré des données personnelles de santé. L'établissement a payé une amende de 2 300 000$ à *HHS*.
University of California, Los Angeles Health	Piratage	4 500 000	2015	Un pirate a accédé à une partie du réseau informatique qui contenait des données personnelles.

TABLE 2.1 – Incidents majeurs des établissements de santé

Les second et troisième plus gros incidents concernent des entreprises gérant plusieurs établissements de soins. Cette gestion particulière est très présente aux Etats-Unis où les établissements sont regroupés.

Les réponses les plus fréquemment citées dans la variable Description sont les suivantes :

Réponse	Nombre
Mise en place de nouvelles technologies de protection	1015
Analyse du risque	1003
Formation des employés	838
Mise en place de nouvelles procédures	659
Prestations de protection contre l'usurpation d'identité et de surveillance du crédit aux individus affecté	388
A bénéficié de l'assistance d'OCR	292
Sanction des employés impliqués	280
Changement des mots de passe	195
Mise en place de moyens de protection physique	192
Adoption de technologies de chiffrage	146
Réalisation d'enquêtes	130
Mise en place de nouveaux plans de gestion du risque	107
Modification du contrat avec le partenaire	53
Fin du contrat avec le partenaire	51
Implémentation d'évaluations périodiques	15

TABLE 2.2 – Réponses les plus fréquentes

Il est intéressant de noter que la réponse la plus fréquente est la mise en place de nouvelles technologies de protection. Cela peut concerner par exemple l'installation d'un nouveau pare-feu. Les établissements conduisent très souvent une analyse du risque après un incident. La troisième réaction la plus fréquence est la formation des employés, ce qui montre que les établissements ont conscience qu'une sensibilisation du personnel à ce risque peut réduire de manière significative les risques informatiques.

Le graphique suivant indique la moyenne d'individus affectés par type d'incident. Il est possible d'observer que ce nombre est beaucoup plus élevé pour les piratages.

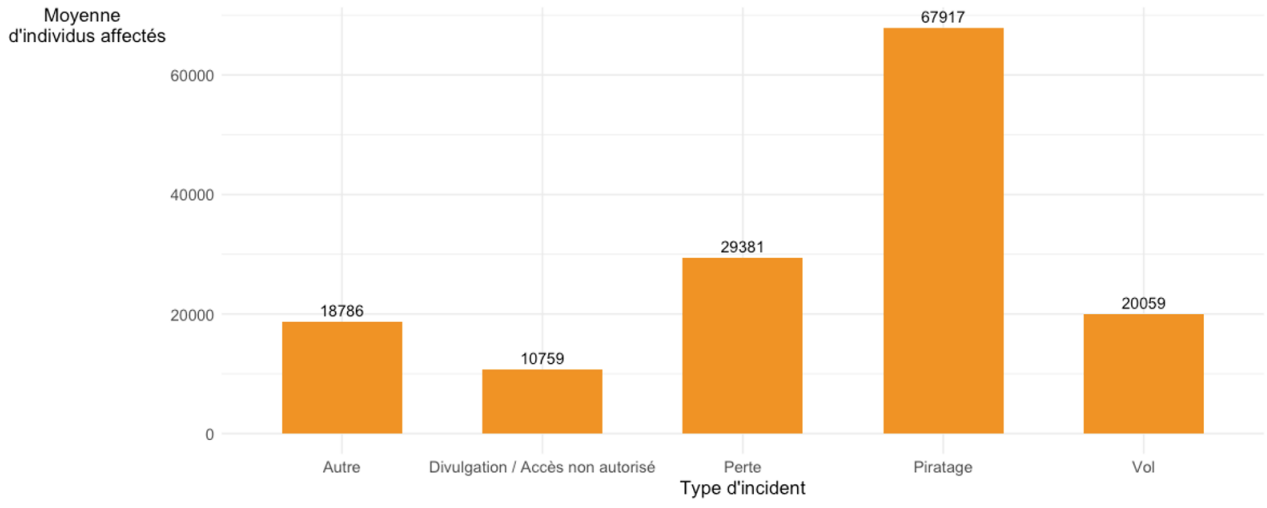


FIGURE 2.3 – Moyenne d'individus affectés par type d'incident

Ensuite, le nombre de mentions de chaque localisation dans la variable Localisation des données est illustré ci-dessous. En effet, un incident peut concerner plusieurs localisations. Les deux localisations les plus fréquentes sont le serveur réseau et l'email.

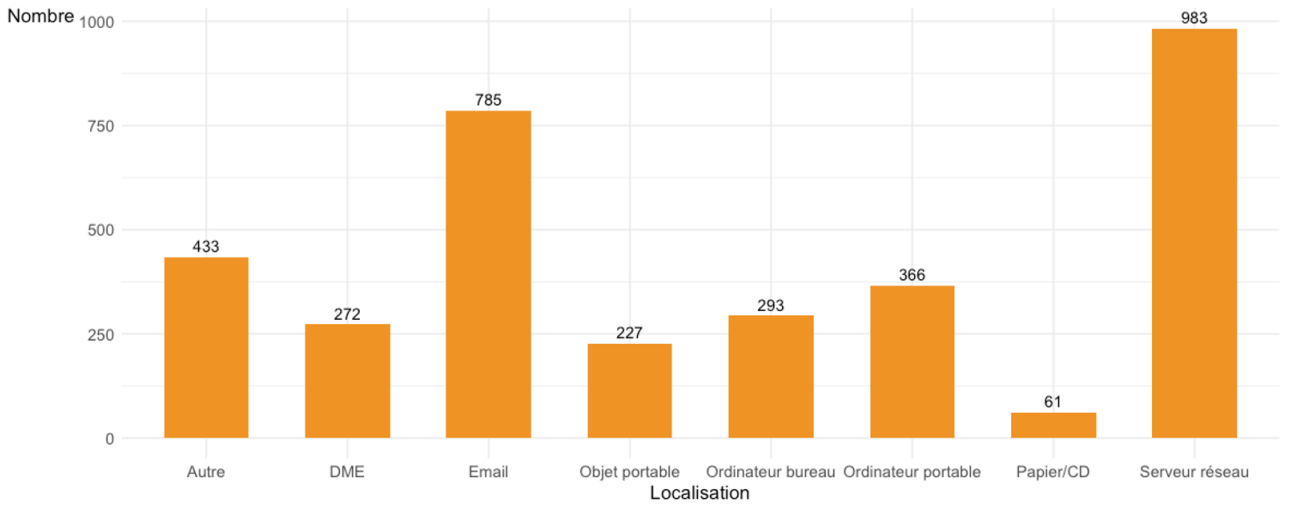


FIGURE 2.4 – Nombre d'incidents par localisation

Le graphique ci-dessous illustre le nombre d'incidents par type de données concernées. Les données démographiques et cliniques sont le plus souvent concernées. Il est très rare que seuls des données financières ou cliniques soient concernées. En effet, les données démographiques constituent un premier niveau d'information, sans lesquelles les données cliniques sont rarement présentes. Les données financières semblent être plus rares et sont presque toujours associées à des données démographiques et cliniques.

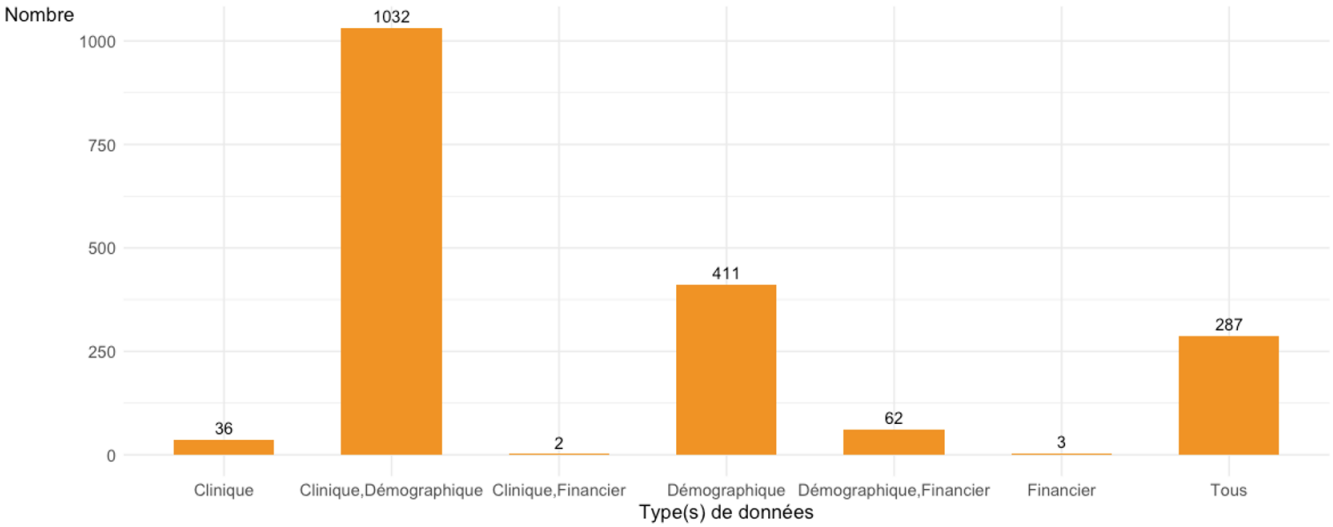


FIGURE 2.5 – Nombre d'incidents pour chaque type de données concernées

2.2.5 Sélection des incidents pour la modélisation

Les incidents déclarés dans la base de données sont variés et ne sont pas tous assimilables à un sinistre Cyber du point de vue de Relyens. La variable description a donc permis de supprimer des incidents qui correspondaient à des cas particuliers d'exclusions mentionnés dans le contrat de Relyens.

Également, les conditions de déclaration des incidents ont été finalisées en 2013. Les incidents de

2014 à 2021 seront donc retenus pour la modélisation.

La première analyse exploratoire a donné des informations sur les incidents Cyber. Afin de pouvoir étudier leur fréquence, il est nécessaire de comparer les établissements qui ont déclaré des incidents avec ceux qui n'en ont pas déclaré à partir de variables explicatives communes. Une autre base est donc recherchée, répertoriant des établissements de santé avec des informations les concernant, dont certains ont déclaré des incidents.

2.3 Présentation de la base de données explicatives utilisée pour l'étude de la fréquence

2.3.1 La base *Hospital Provider Cost Report 2018* comme base contrats

Afin de mesurer la fréquence et l'intensité des incidents selon les caractéristiques des établissements, une autre base est utilisée et joue le rôle de « base contrats » : chaque établissement est associé à sa ou ses déclarations de violations de données s'il est concerné.

Les données de *Hospital Provider Cost Report 2018* sont principalement des informations financières ou structurelles déclarées par les hôpitaux à *Medicare* ou *Medicaid* en 2018. Elle recense 6045 hôpitaux ce qui est proche du nombre d'hôpitaux annuel aux Etats-Unis. En effet *American Hospital Association (AHA)* dénombre 6146 hôpitaux à partir des données fiscales de 2018. [28]. Les hôpitaux absents de cette base ne sont donc pas étudiés pour conserver une vision annuelle. Par exemple, les hôpitaux pour anciens combattants sont absents des données, ainsi que ceux ayant fermé avant 2018 ou ouvert après 2018.

Lors de la rédaction de ce mémoire, 2018 était l'année la plus récente pour laquelle des données étaient disponibles sous forme de base, appelée *Public Use File*. Elles sont jugées assez récentes pour l'étude, car il y a généralement peu de changements dans la structure des établissements d'une année à l'autre. Dans la suite du mémoire, *Hospital Provider Cost Report 2018* est nommée la base contrats. D'autres bases peuvent être trouvées portant sur d'autres catégories d'établissements, comme les maisons de soins médicalisées [29] mais peu de correspondances avec les incidents ont été trouvées, ou encore les groupes de médecins [30] où un certain nombre de correspondances ont été trouvées mais la base indique une unique variable explicative.

2.3.2 Méthode pour identifier et associer les incidents Cyber aux établissements médicaux de la base contrats les ayant déclarés

Chaque établissement de la base contrats est recherché au sein de la base des incidents de violations de données personnelles afin de l'associer à ses éventuelles déclarations. Pour cela, les noms d'établissements renseignés dans les bases sont comparés par différentes méthodes de correspondances approximatives. La variable Etat est également utilisée pour ne rechercher des correspondances que si les établissements se trouvent dans le même Etat.

Dans un premier temps, l'uniformisation des noms d'établissements des deux bases a été réalisée par exemple en mettant en minuscule, en supprimant les ponctuations, ou en développant les abréviations

28. *AHA Hospital Statistics, 2020 edition*
29. Base disponible sur le site : https://data.cms.gov/provider-compliance/cost-report/skilled-nursing-facility-cost-report
30. *Group Practice Linkage File* disponible sur https://www.ahrq.gov/chsp/data-resources/compendium-2018.html

les plus fréquentes.

Chaque base est ensuite divisée selon la variable Etat et les correspondances sont recherchées uniquement entre les établissements se trouvant dans le même Etat, ce qui limite le nombre de fausses correspondances et permet de gagner en efficacité informatique.

Plusieurs fonctions sont implémentées sous le logiciel RStudio® pour rechercher des ressemblances entre les noms. Chaque fonction crée une correspondance entre une ligne de la base contrats et une ligne de la base des incidents si la ressemblance est suffisamment forte selon un critère. Concrètement, pour chaque méthode, une colonne est créée dans la base des incidents, et si une correspondance est trouvée, le numéro de ligne correspondant de la base contrats y est renseigné.

Les fonctions dépendent chacune d'un des critères suivants :

— Présence du nom de l'établissement au sein d'un nom de l'autre base, en autorisant les ajouts de caractères avant et après
— Présence de tous les mots du nom de l'établissement au sein d'un nom de l'autre base, sans regarder l'ordre
— Distance maximum entre un nom de la base contrats et un nom de la base sinistres selon différentes méthodes de calcul détaillées en annexe (Distance de Damerau-Levenshtein, distance de Jaro-Winkler, distance n-gramme, sous séquence commune la plus longue). La distance maximum choisie est de 20% multiplié par le nombre de caractères du nom indiqué dans la base des incidents. Cette limite a été choisie arbitrairement après plusieurs essais car elle semble être optimale pour permettre un certain nombre de correspondances sans pour autant créer des liens entre des noms très éloignés.

Lorsqu'une fonction retient plusieurs correspondances possibles avec la base contrats, aucun résultat n'est retenu par la fonction car la base contrats ne contient pas de doublons après retraitement, et il n'est pas possible de déterminer automatiquement la bonne correspondance. Ainsi, pour maximiser les chances de création de correspondances, les fonctions ont également été appliquées sur les noms des établissements combinés au nom de leur ville, du canton, et sur les noms d'établissements dont les mots sont triés par ordre alphabétique. Pour les incidents sans correspondances, les fonctions sont aussi appliquées aux éventuels anciens noms d'établissements. Ces anciens noms sont extraits à partir des autres bases nommées *Cost Report* publiées avant 2018 par *CMS* et conservés s'ils diffèrent du nom en 2018. Chaque résultat a été ensuite vérifié visuellement, ce qui était possible compte tenu du faible nombre de données. Les mauvaises correspondances sont manuellement enlevées. Cela est nécessaire car les noms d'établissements peuvent parfois être très proches mais ne pas correspondre au même établissement. Il a également été vérifié manuellement que tous les hôpitaux de la base contrats ayant déclaré un incident ont été trouvés par les fonctions, en recherchant les mots clés *Hospital* et *Medical center* dans les noms de la base sinistres sans correspondances.

Pour trois incidents, deux établissements sont cités dans la variable Nom de l'établissement. Le retraitement effectué est le même que celui présenté dans le mémoire *Modélisation assurantielle du risque cyber* d'Anaïs Martinez qui étudie notamment la base de *Privacy Rights Clearinghouse*. Du point de vue de l'assureur, deux contrats Cyber peuvent être déclenchés dans ce cas. Les lignes concernant ces incidents sont donc dupliquées. L'hypothèse est faite que le nombre d'individus affectés de chaque établissement est la moitié du nombre d'individus touchés par l'incident. Cela permet d'affecter à chacun des établissements le sinistre.

Ainsi, 568 des 6045 hôpitaux de la base contrats a déclaré un incident. Après sélection des incidents pouvant être considérés comme un incident Cyber d'après le contrat proposé par Relyens et sélection des années après 2013, 415 incidents sont retenus.

Il est possible de voir en recherchant le mot clé *Hospital* que 15 hôpitaux n'ont pas été trouvé dans la base contrats, pour des raisons variées ou inconnues. Par exemple, l'hôpital *Integrity Transitional Hospital* a fermé ses portes, ou bien *Martin Army Community Hospital* est un hôpital militaire qui ne fait pas partie du périmètre de la base contrats.

2.3.3 Focalisation sur les variables disponibles sur les hôpitaux

Les données *Cost Report* publiées par *The Centers for Medicare and Medicaid Services (CMS)* apportent différentes informations sur des établissements de santé comme le nombre de lits et d'employés, leurs actifs, leurs coûts ainsi que leurs revenus. En particulier, *Hospital Cost Report 2018 Public Use File* est la base la plus récente, regroupant les mesures les plus communes des déclarations des hôpitaux pour les périodes fiscales couvrant 2018. Elle permet d'étudier et de sélectionner des variables explicatives pour le modèle.

D'autres informations sont accessibles mais ne sont pas directement exploitables car cela nécessite de construire une base en sélectionnant chaque variable d'intérêt parmi une quantité importante d'informations disponibles. En effet, les hôpitaux doivent remplir des formulaires contenant de nombreuses informations financières détaillées. Des informations plus récentes, d'autres variables ou encore des informations similaires pour d'autres types d'établissements peuvent donc être extraites depuis ces formulaires grâce au package R *medicare*, en indiquant pour chaque information souhaitée son numéro correspondant dans le formulaire. Construire une base avec cette méthode serait coûteux en temps.

Ainsi, seule la base *Hospital Cost Report 2018 Public Use File* est étudiée, à laquelle sont ajoutées les variables binaires suivantes, vues dans les formulaires : Certification données, qui est une certification pour l'utilisation des données électroniques de santé [31], Accueille des internes, Hôpital d'accès critique [32]. Au moment de la rédaction de ce mémoire, la variable Type de soins semblait erronée dans la base *Public Use File* car les modalités ne correspondaient pas aux établissements associés, elle a donc été également extraite du formulaire.

Les données de la base retenue s'étendent de 2017 à 2019, ce qui est jugé assez récent. Cela permet d'évaluer la structure globale des hôpitaux et de ne pas prendre en compte les éventuelles perturbations dues à la crise COVID. Ces données explicatives sont associées aux incidents de données déclarées entre 2014 et 2021. L'hypothèse faite est que la structure des hôpitaux varie généralement peu entre les années, rechercher les données plus anciennes dans le but d'associer à chaque établissement ses données au moment de l'incident aurait été coûteux en temps et apporterait un gain faible dans l'objectif de ce mémoire.

Les grandes catégories des variables présentes sont :
- Informations sur les établissements (nom, numéro d'identification CCN, adresse, Etat, catégorie,...)
- Informations sur la taille de l'hôpital (nombre de lits, nombre d'employés Equivalent Temps Plein, nombre de séjours)
- Différentes catégories de coûts et dépenses de l'établissement
- Valeur de différents actifs courants, immobilisations, et passifs
- Revenus pour plusieurs catégories d'activité

31. Le terme *Meaningful use* est utilisé par *CMS* pour identifier les établissements respectant certains standards de qualité et de sécurité dans l'utilisation des données personnelles de santé électroniques, et sont par conséquent éligibles à un programme de financement

32. Les hôpitaux d'accès critique sont des hôpitaux issus d'un programme qui vise à développer la présence hospitalière dans certaines zones

— Résultat net pour plusieurs catégories d'activité

Face au nombre important de variables, certaines d'entre elles sont présélectionnées : il s'agit des variables disposant de moins de 30% de valeurs manquantes et indiquant une information globale sur l'établissement, par exemple qui concernent toutes les catégories d'activité. Un tableau en annexe E décrit les variables retenues après ces sélections ainsi que leur pourcentage de valeurs manquantes.

De plus, certains établissements apparaissent plusieurs fois dans la base lorsqu'ils ont réalisé plusieurs déclarations à des périodes fiscales différentes, par exemple en milieu d'année puis fin d'année. Les doublons sont supprimés en conservant en priorité la période fiscale d'un an, puis la plus longue. Après ce traitement, près de 96% des déclarations correspondent à une période fiscale de 364 jours soit un an. Les variables qui dépendent de la durée de cette période fiscale sont ajustées pour que toutes les valeurs correspondent à une période fiscale d'un an. Leurs valeurs sont donc multipliées par 364 et divisées par la durée de la période fiscale en jours.

La base *Hospital General Information* de 2019 [33] publiée par *CMS* est également utilisée pour ajouter la variable Service d'urgences, qui désigne la présence de cette activité ou non.

2.3.4 Gestion des valeurs manquantes

Des valeurs manquantes sont observées à la fois pour des variables qualitatives et quantitatives. Une première imputation des valeurs manquantes catégorielles est réalisée à partir d'autres sources de données. En effet, ces variables sont peu susceptibles de varier selon les années et sources de publication : il s'agit des variables Accueille des internes, Hôpital d'accès critique, Type de soins, et Certification données, extraites des formulaires *Cost Report* publiés par *CMS*. Les valeurs manquantes de ces variables sont complétées par les valeurs indiquées dans les formulaires de 2019, puis de 2017 pour les valeurs manquantes restantes.

Après ces recherches, des valeurs manquantes sont encore présentes. Deux méthodes permettant l'imputation à la fois de variables quantitatives et qualitatives sont testées. Il s'agit des k plus proches voisins et de la forêt aléatoire. Pour des questions d'efficacité, la base utilisée pour l'imputation des valeurs manquantes est celle des variables présélectionnées, qui sont listées en annexe E. En effet, l'objectif est d'imputer les variables d'intérêt pour le modèle, à partir des autres variables. Le nombre de variables est important, et certaines sont très peu complétées. Un compromis a donc dû être trouvé pour veiller à garder suffisamment de variables pour prédire au mieux les valeurs manquantes, tout en n'alourdissant pas le code par des variables superflues. En particulier les variables avec beaucoup de valeurs manquantes ne sont pas retenues pour la méthode.

La première méthode utilisée est l'algorithme des K plus proches voisins. Il identifie les k voisins les plus proches de la ligne où se trouve la donnée manquante selon la mesure de distance de Gower qui s'applique aux données numériques et catégorielles. Pour p variables et deux observations i et j, cette distance est définie par :

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \sigma_{i,j,k}}{\sum_{k=1}^{p} w_k}$$

avec w_k le poids de la k-ième variable, par défaut 1. Puis

$$\sigma_{i,j,k} = |x_{i,k} - x_{j,k}|/r_k$$

où $x_{i,k}$ est la valeur de la k-ième variable de l'observation i et r_k la différence entre la valeur maximale et minimale de la variable k.

33. Base disponible sur le site : https://data.cms.gov/provider-data/archived-data/hospitals

Pour les variables catégorielles,

$$\sigma_{i,j,k} = \begin{cases} 0 & \text{si } x_{i,k} = x_{j,k} \\ 1 & \text{si } x_{i,k} \neq x_{j,k} \end{cases}$$

L'algorithme impute ensuite aux valeurs manquantes la médiane des valeurs observées ou la catégorie la plus fréquente chez les k plus proches voisins.

La seconde méthode est l'imputation par une forêt aléatoire avec le package *missForest* du langage R. Il se base sur les forêts aléatoires de Breiman qui utilisent les arbres de décisions et l'agrégation de modèles.

Un arbre est constitué d'une racine qui correspond à toutes les données, de branches qui divisent les données et de feuilles qui sont des sous groupes. Chaque noeud divise les données en deux selon une règle. Chaque règle est testée pour maximiser l'homogénéité des groupes finaux. Un certain nombre d'arbres sont générés, chacun construit sur un sous-ensemble aléatoire des données, et le résultat sera une moyenne de ces arbres, ce qui assure une stabilité. Les étapes de l'algorithme sont les suivantes [34] :

— Première imputation par la moyenne
— Classement des variables par ordre croissant de part de valeurs manquantes.
— Pour chaque variable, les données précédemment imputées sont remplacées par des valeurs manquantes uniquement pour cette variable et sont imputées par une forêt aléatoire à partir des autres variables.
— La différence entre les valeurs imputées à l'étape actuelle et celles à l'étape précédente est calculée. L'étape précédente est réitérée, c'est à dire que les imputations sont à nouveau réalisées et leur différence calculée, tant que cette différence diminue.

L'algorithme fournit également l'erreur *OOB* (*out of bag*). Elle est calculée à partir des données observables. Pour chaque donnée observable, l'algorithme construit des arbres sans l'information de cette donnée et impute sa valeur. La différence entre chaque donnée observée et imputée est calculée, puis l'erreur *OOB* est estimée par la moyenne de ces différences.

Les deux critères à minimiser choisis pour la détermination des paramètres des méthodes et pour la sélection de la méthode la plus performante sont les suivants :

Pour les variables quantitatives, la racine de l'erreur quadratique moyenne (NRMSE) est définie par :

$$NRMSE = \sqrt{\frac{Moy\left((X_{obs} - X_{imp})^2\right)}{Var\left(X_{obs}\right)}}$$

avec X_{obs} le vecteur des valeurs observées, X_{imp} celui avec les valeurs imputées, Moy la moyenne empirique et Var la variance empirique. Pour les variables qualitatives, la proportion de données mal classées (PFC) est définie par :

$$PFC = \frac{\text{Nombre de données mal classées}}{\text{Nombre de données classées}}$$

Les indicateurs sont calculés sur l'ensemble des données imputées.

34. DIXNEUF P. Analyse de la performance de la méthode d'imputation de données manquantes missforest et application à des données environnementales. Mémoire de maîtrise électronique, Montréal, Ecole de technologie supérieure, 2019

Afin de pouvoir comparer les deux méthodes, elles sont appliquées sur la base contrats sans les lignes contenant des valeurs manquantes, qui compte après ce traitement 3324 lignes. Les variables Nom, Ville, Code Postal, Numéro d'identification CCN et Numéro de déclaration sont retirées. Des valeurs manquantes sont créées au sein de cette base dans les mêmes proportions que celles des réelles données manquantes de la base contrats. Le NRMSE et PFC sont calculés et comparés pour toutes les données imputées. Pour l'imputation par forêt aléatoire, les variables qualitatives Etat et Canton sont converties en variables binaires pour chaque modalité car l'algorithme ne permettait pas l'utilisation d'une variable quantitative avec un nombre important de modalités. Dans un souci de comparaison, les variables sont reconverties en variables qualitatives avant calcul du NRMSE et PFC.

La méthode des K plus proches voisins est appliquée pour différentes valeurs de k entre 5 et 50, et celle de la forêt aléatoire pour un nombre d'arbres allant de 20 à 120 arbres. Les meilleurs résultats obtenus par chaque méthode sont résumés dans le tableau suivant :

	K plus proches voisins K=5	Forêt Aléatoire 30 arbres
NRMSE	0,49	0,33
PFC	0,04	0,02

Ainsi, la méthode de la forêt aléatoire est plus performante et sera utilisée pour estimer les valeurs manquantes.

Cette méthode est donc retenue et appliquée sur la base contrats. L'erreur *out of bag* est de 0,23 pour le NRMSE et 0,03 pour le PFC.

2.3.5 Enrichissement de la base contrats par des nouvelles variables

Afin de mesurer au mieux la santé financière des établissements, de nouvelles variables sont calculées à partir de celles déjà présentes dans la base, par exemple en effectuant des ratios. Ainsi, la variable Chiffre d'affaires sera calculée par :

$$\text{Chiffre d'affaires} = \text{Revenu des soins} + \text{Autres recettes}$$

Les variables construites par ratio sont listées dans le tableau ci-dessous.

Nom	Calcul
Dépenses par lit	$\frac{\text{Dépenses opérationnelles}}{\text{Nombre de lits}}$
Durée moyenne séjour	$\frac{\text{Nombre de journées}}{\text{Nombre de sorties}}$
Équipement par lit	$\frac{\text{Valeur équipement mobile}}{\text{Nombre de lits}}$
Employés par journée	$\frac{\text{Nombre d'employés ETP}}{\text{Nombre de journées}}$
Employés par lit	$\frac{\text{Nombre d'employés ETP}}{\text{Nombre de lits}}$
Part de marché	$\frac{\text{Revenu des soins + Autres recettes}}{\text{Revenus de tous les hôpitaux du même Etat}}$
Ratio CA Sorties	$\frac{\text{Revenu des soins + Autres recettes}}{\text{Nombre sorties}}$
Ratio impayés coûts	$\frac{\text{Impayés}}{\text{Coût total}}$
Ratio Recettes Dépenses	$\frac{\text{Revenu soins + Autres recettes}}{\text{Dépenses opérationnelles + Autres dépenses}}$
Revenu par lit	$\frac{\text{Revenu des soins + Autres recettes}}{\text{Nombre de lits}}$
Revenu soins par lit	$\frac{\text{Revenu des soins}}{\text{Nombre de lits}}$
Salaire moyen	$\frac{\text{Montant salaires}}{\text{Nombre d'employés ETP}}$
Sorties par lit	$\frac{\text{Nombre de sorties}}{\text{Nombre de lits}}$
Taux d'occupation	$\frac{\text{Nombre de journées}}{\text{Nombre de jours-lits}}$
Taux de marge globale	$\frac{\text{Résultat net}}{\text{Revenus soins + Autres recettes}}$

TABLE 2.3 – Variables construites par ratios

De plus, deux autres bases sont également utilisées pour identifier des liens de connexion entre les hôpitaux et d'autres établissements.

Compendium of U.S. Health Systems est une base de données qui regroupe 637 systèmes de santé américains [35]. Les systèmes correspondent à une organisation d'établissements de santé qui sont liés au travers d'une propriété ou d'une gestion commune. Les systèmes identifiés contiennent au moins un hôpital et un groupe de praticiens.

La base *EHR (Electronic Health Record) Products Used for Meaningful Use Attestation* publiée par l'*ONC*(*Office of the National Coordinator for Health Information Technology*) [36] liste les noms des logiciels de gestion des données électroniques de santé utilisés par chaque établissement lorsqu'ils remplissent les critères d'utilisation dits *meaningful use* définis par CMS. Les noms des fournisseurs de ces logiciels sont également indiqués.

Ainsi les variables créées sont les suivantes :

— Membre d'un système, avec trois modalités qui sont non membre d'un système, seul hôpital membre du système, et membre du système avec d'autres hôpitaux
— Logiciel externe données, qui désigne l'utilisation d'un logiciel de gestion de données de santé certifié, dont la modalité est oui ou non.

2.3.6 Statistiques descriptives

Une première analyse de la base contrats est réalisée afin d'étudier les profils d'hôpitaux présents dans les données. Les périodes fiscales présentes sont étalées sur les années 2017-2018, l'année 2018 ou bien la période 2018-2019. Les statistiques suivantes décrivent les établissements.

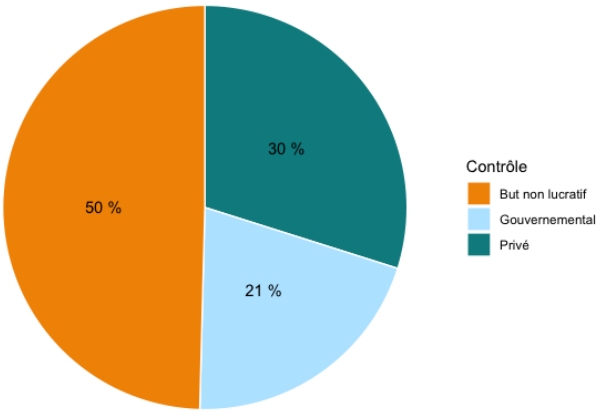


FIGURE 2.6 – Part d'établissements selon le type de contrôle

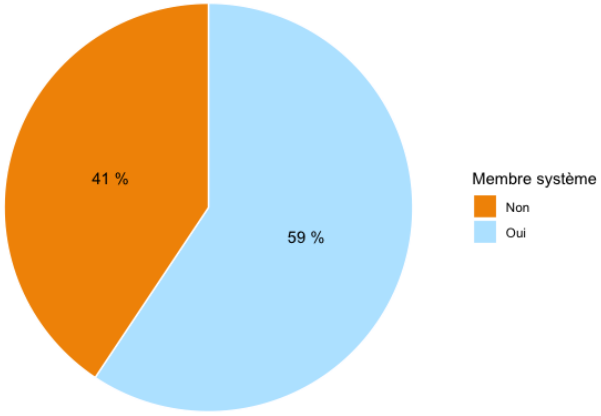


FIGURE 2.7 – Part d'établissements membres d'un système

Une majorité d'hôpitaux est à but non lucratif, et membre d'un système. Le graphique ci-dessous illustre le nombre d'hôpitaux par type de soins prodigués. Les soins de court terme sont très largement majoritaires.

35. Base disponible sur le site : https://www.ahrq.gov/chsp/data-resources/compendium-2018.html
36. Base disponible sur le site : https://www.healthit.gov/data/datasets/ehr-products-used-meaningful-use-attestation

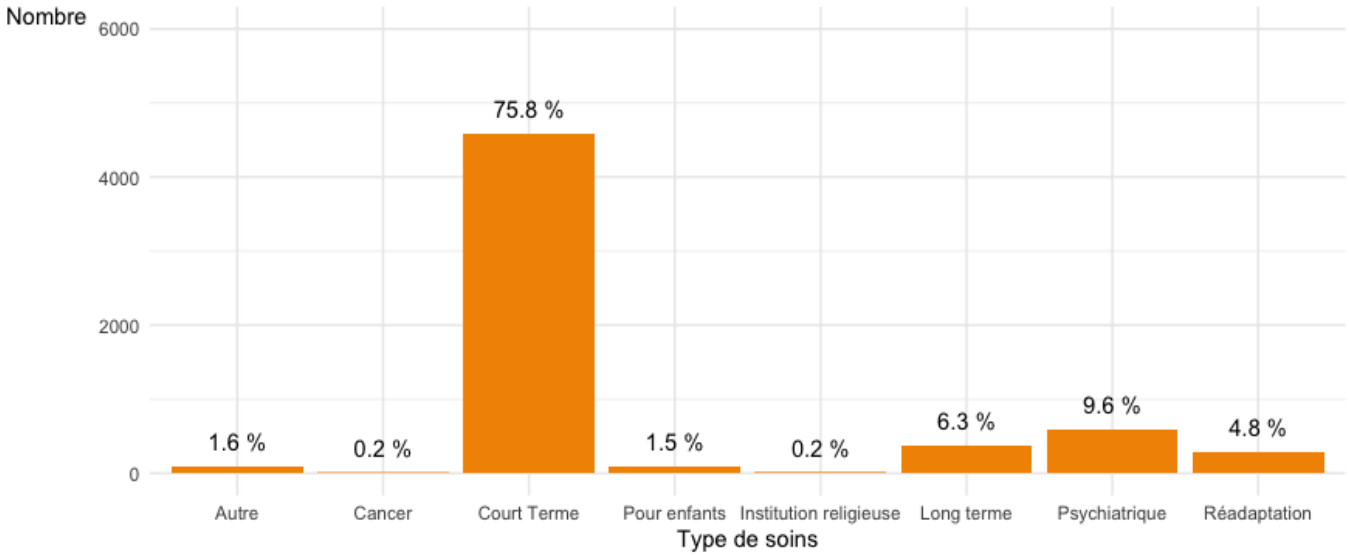


FIGURE 2.8 – Nombre d'hôpitaux par type de soins

Il est également intéressant de noter que 96% des données correspondent à une période fiscale d'un an. Pour les autres, un ajustement proportionnel est effectué afin d'estimer pour les variables dépendant du temps une valeur annuelle.

Pour chaque variable, la distribution des valeurs, notamment celles aberrantes, est observée grâce à une boîte à moustaches. Les variables quantitatives sont toutes très dispersées. Par exemple, les boîtes à moustaches des variables Nombre d'employés ETP, Nombre de lits et Revenus hospitalisations permettent de visualiser une forte dispersion :

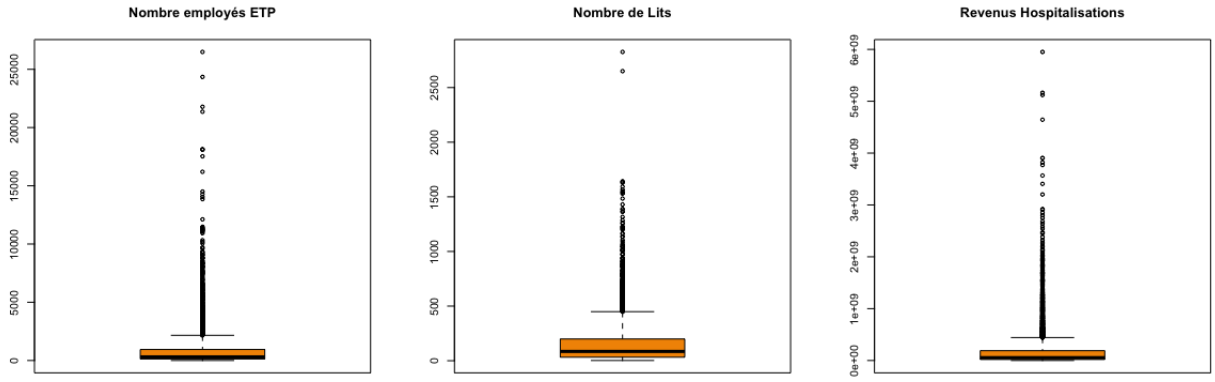


FIGURE 2.9 – Boîte à moustaches des variables Nombre d'employés ETP, Nombre de lits, et Revenus Hospitalisations

Il est difficile d'identifier les valeurs aberrantes parmi les nombreuses valeurs très faibles ou très élevées. Par exemple, un nombre de lits très faible peut paraître aberrant, mais cela est possible lorsque l'activité principale correspond aux consultations et non à des hospitalisations. Par exemple, l'hôpital *Cass Lake* indique sur son site qu'il a seulement 5 lits. [37]

37. Source : https://www.ihs.gov/bemidji/healthcarefacilities/casslake/

Face au peu de données d'incidents réalisés, il est choisi de ne pas supprimer les lignes contenant des valeurs aberrantes. Également, le nombre de variables étant important, une méthode de correction des valeurs aberrantes efficace doit être mise en place. Un seuil minimal et maximal de valeur possible est appliqué pour chaque variable qui est utilisée dans un modèle. Cela est réalisé après le calcul des variables par ratio. En effet, après calcul des ratios, des variables peuvent se révéler aberrantes alors que le numérateur et le dénominateur n'ont pas de valeurs aberrantes, lorsque les deux variables ne sont pas cohérentes entre elles. De même, une correction avant calcul des ratios pourrait aboutir à des erreurs, dans le cas ou seule une des deux variables servant au ratio serait corrigée.

Ainsi, le tableau suivant résume pour chaque variable concernée le nombre de valeurs qui ont été remplacées par un seuil minimal ou maximal, lorsqu'elles dépassaient ce seuil, déterminé graphiquement par analyse des boîtes à moustaches.

Variable	Nombre de valeurs aberrantes	Seuil
Autres recettes	145	Minimum 0
Dépenses par lit	12	Maximum 10 000 000
Durée moyenne séjour	101	Maximum 400
Équipement par lit	1	Maximum 20 000 000
Nombre d'employés ETP	10	Minimum 5
Nombre d'employés ETP par lit	30	Maximum 30
Ratio coûts Charges	141	Maximum 5
Ratio Impayés Couts	8	Maximum 1
Ratio Recettes Depenses	5	Maximum 4
Revenu soins	11	Minimum 0
Salaire moyen	9	Minimum 10 000
Salaire moyen	36	Maximum 300 000
Sorties par lit	14	Maximum 100
Taux d'occupation	1274	Maximum 1
Taux de marge globale	10	Minimum -2

Les corrections concernent toutes moins de 150 valeurs des variables et auront donc un effet limité sur la modélisation, mis à part le Taux d'occupation. De nombreuses valeurs sont supérieures à 1 ce qui est dû au fait que les variables considérées sont celles de toutes les activités de soins, et que des journées sont probablement comptabilisées alors qu'elles n'ont pas nécessité l'utilisation d'un lit et ne sont pas comptabilisées dans les jours lits. Ce retraitement a tout de même été effectué car il est souhaité pour cette variable de regrouper ensemble toutes les valeurs supérieures à 1. Pour Autres recettes et Revenu soins, le retraitement a eu lieu avant le calcul des variables ratios pour éviter les incohérences dans les calculs des ratios liées à des revenus qui pouvaient être négatifs.

En résumé, les déclarations de violations de données personnelles de santé à *HHS OCR* semblent être les données publiques les plus adaptées dans le contexte de ce mémoire. En effet, elles sont régulièrement alimentées, et ciblent les établissements de santé. De plus, le caractère obligatoire de la déclaration rend possible une modélisation de la fréquence. Il faut cependant noter que la base ne contient pas d'information sur le coût des incidents, mais sur le nombre d'individus affectés. Également, elle concerne un périmètre non exhaustif des incidents Cyber, qui sont les incidents touchant des données personnelles de santé au dessus d'un certain seuil d'individus affectés. Une sélection des incidents est réalisée pour se rapprocher du contexte de l'assurance commercialisée par le groupe Relyens, par exemple l'exclusion des incidents de données sur papier qui ne sont pas des incidents Cyber.

L'étude de ces données révèle une très forte hausse des piratages au cours des années. Une proportion non négligeable de piratages est associée à un incident d'un prestataire de l'établissement. De plus, la moyenne des individus affectés varie selon la catégorie d'incident, elle est plus élevée pour les piratages. Les variables créées à partir de la variable Description mettent en avant que la majorité des incidents concernent des données démographiques et cliniques, et que la réponse à l'incident la plus fréquente est la mise en place de nouvelles technologies de protection.

Afin d'étudier la fréquence, la base *Hospital Cost Report* de *CMS* est choisie pour jouer le rôle d'une base contrats. Les éventuelles déclarations de chaque hôpital sont recherchées en comparant les noms d'établissements et les Etats fédérés. Pour cela, différentes méthodes pour identifier deux noms d'établissements proches entre les deux bases sont appliquées. La base contrats regroupe des informations financières et structurelles sur 6 045 hôpitaux américains, ce qui est proche du nombre annuel d'hôpitaux aux Etats-Unis.

La base contrats contient beaucoup de variables, seules les variables apportant une information générale et ne contenant pas trop de valeurs manquantes sont retenues. Les valeurs manquantes de la base contrats sont ensuite imputées. Deux méthodes sont comparées : les K plus proches voisins, puis la forêt aléatoire qui se révèle être plus performante et est donc retenue. Des variables sont construites en effectuant le ratio entre des variables existantes afin d'ajouter une information supplémentaire. Enfin, une correction de valeurs aberrantes est effectuée en appliquant une valeur minimum et maximum à certaines variables.

Chapitre 3

Modélisation du risque de violation de données personnelles de santé pour les hôpitaux américains

La base de données choisie et retraitée adopte le rôle d'une base contrats classique en assurance. Elle permet ainsi d'associer à chaque établissement un ou plusieurs éventuels sinistres, et de modéliser une fréquence ou une probabilité de sinistre. L'approche de tarification appliquée dans ce mémoire est présentée, avant d'être complétée par des rappels théoriques. Pour chaque catégorie d'incident, s'il y a assez de données, une modélisation de la probabilité de survenance d'un sinistre et du nombre d'individus affectés pour les hôpitaux américains sera effectuée. Dans le cas contraire, la moyenne empirique est retenue. L'ensemble des informations disponibles dans les bases américaines seront utilisées dans ce chapitre afin d'étudier leurs effets, même si elles ne sont pas toutes comparables à des informations existantes sur les hôpitaux en France.

3.1 Approche de tarification retenue

Les données utilisées dans le cadre de ce mémoire ne sont pas des données sinistres d'assurance. En effet, le coût des incidents est inconnu, ce qui ne permet pas d'appliquer directement les méthodes de tarification traditionnelles en assurance pour modéliser le coût. De plus, le faible volume des données doit amener à une interprétation prudente des résultats. Il sera choisi de construire un modèle de la forme fréquence x coût en effectuant des estimations lorsque les informations ne sont pas disponibles.

Dans un premier temps, pour chaque catégorie d'incident avec suffisamment de données, un modèle d'estimation de la fréquence sera construit. Il sera vu que peu d'établissements ont déclaré plus d'un incident du même type. Ainsi, une régression logistique sera estimée pour modéliser la probabilité de survenance de chaque type d'incident, avec des méthodes pour répondre au déséquilibre des classes. Ce modèle sera calibré sur l'ensemble des données entre 2014 et 2021, puis la probabilité estimée sera ajustée afin d'aboutir à une probabilité annualisée. Elle sera ensuite également ajustée pour correspondre à une fréquence dans un contexte du secteur de la santé en France. La probabilité des incidents avec peu de données sera estimée par moyenne empirique et fera également l'objet des mêmes ajustements.

Dans un second temps, il s'agit de déterminer une estimation des coûts pour les multiplier aux

fréquences estimées. L'Institut Ponemon est un institut de recherche qui effectue des études sur la sécurité des données informatiques. Il publie avec la société IBM des études qui estiment un coût pour un enregistrement compromis. Le montant d'un sinistre pourrait donc être estimé par le nombre d'individus affectés multiplié par ce coût. Cette approche a été utilisée dans le mémoire de Florian Pons [38]. Elle n'est pas favorisée dans le contexte présent car l'intensité d'un sinistre n'est pas directement liée au nombre d'individus affectés dans le secteur de la santé et explique seulement une partie du coût du sinistre. Par exemple, une attaque Cyber qui perturbe le déroulement des soins d'un établissement pendant une longue durée peut avoir un préjudice financier plus important qu'une divulgation de données de patients sans impact sur le fonctionnement d'un établissement. Également, le coût présenté pour le secteur de la santé dans l'étude de 2021 [39] est mondial et est probablement tiré vers le haut par les Etats-Unis. En effet, les amendes infligées aux établissements de santé américains ne respectant pas la confidentialité des données sont dans les faits beaucoup plus élevées que celles en France.

Afin d'être au plus proche de la réalité, il a été choisi de profiter de l'expertise des experts en risque informatique de l'entreprise. Ainsi, pour chaque garantie du contrat d'assurance proposé par Relyens, un coût est déterminé par avis d'expert, pouvant varier selon différents éléments comme la catégorie d'incident, les caractéristiques des établissements, et dans certains cas le nombre d'individus affectés. Une probabilité de déclenchement de chaque garantie est aussi déterminée selon le type d'incident. En effet, elles ne sont pas toutes forcément activées après un sinistre, les conséquences d'un incident Cyber pouvant être variées.

Le schéma suivant résume l'approche :

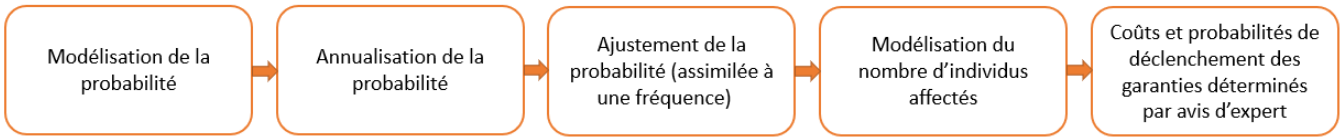


FIGURE 3.1 – Résumé des étapes du modèle de tarification

Cette approche peut se résumer sous forme d'équations. Soit N_i la variable aléatoire qui correspond à la fréquence d'un incident Cyber pendant l'année pour une catégorie d'incident. A_i, B_i, et C_i sont trois évènements qui correspondent au déclenchement d'une garantie pour l'assuré i sachant qu'il a eu un incident d'une certaine catégorie. Les probabilités estimées respectives des évènements pour chaque assuré sont p_a, p_b et p_c, et leur coût respectif sont a, b et c. Y_i est la variable aléatoire du coût pour l'assuré i, et P_i la variable aléatoire du nombre d'individus affectés. Il est considéré dans l'exemple que c doit être multiplié par le nombre d'individus affectés.

Grâce à l'expérience des experts en risque Cyber de l'entreprise, les coûts a, b et c et les probabilités sont définis pour chaque garantie selon la catégorie d'incident et peuvent dépendre des caractéristiques des établissements, et du nombre d'individus affectés. Le coût global d'un incident s'exprime de la manière suivante :

$$Y_i = a\mathbb{1}_{A_i} + b\mathbb{1}_{B_i} + c\mathbb{1}_{C_i}P_i$$

Et

38. PONS F., Etude Actuarielle du Cyber-Risque. *Institut des Actuaires*, 2014
39. *Cost of a Data Breach Report 2021*,IBM

$$\mathbb{E}[Y_i|X_i] = \mathbb{E}[a|X_i] \times p_a + \mathbb{E}[b|X_i] \times p_b + \mathbb{E}[c|X_i] \times p_c \times \mathbb{E}[P_i|X_i]$$

Puis, en considérant que la fréquence et le coût sont indépendants, la prime pure pour un incident vaut selon le modèle fréquence × coût :

$$\mathbb{E}[S_i|X_i] = \mathbb{E}[N_i|X_i] \times \mathbb{E}[Y_i|X_i]$$

Il suffit ensuite de sommer les primes pures de chaque type d'incident qui sont considérés comme indépendants.

3.2 Tarification en assurance

Soit S_i la variable aléatoire représentant la perte pour un contrat i. La tarification repose sur le principe de la mutualisation qui provient de la loi forte des grands nombres. D'après cette loi, la moyenne empirique des pertes converge vers l'espérance de la loi des pertes. Autrement dit, pour un portefeuille de n contrats avec $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$,

$$\overline{S} \underset{n\to+\infty}{\xrightarrow{p.s}} \mathbb{E}[S_i]$$

Dans l'approche fréquence × coût, S_i est exprimé selon deux variables aléatoires.

$$S_i = \sum_{k=1}^{N_i} Y_{i,k}$$

où N_i est la variable aléatoire du nombre de sinistres du contrat i, et $Y_{i,k}$ la variable aléatoire du montant du sinistre k de l'assuré i. Les hypothèses nécessaires sont l'indépendance de N_i et $Y_{i,k}$, et les variables aléatoires $Y_{i,k}$ doivent être indépendantes et identiquement distribuées.

Ainsi, $\mathbb{E}[S_i] = \mathbb{E}[N_i]\mathbb{E}[Y_{i,k}]$ désigne la prime pure. En prenant en compte les variables explicatives, la prime pure devient $\mathbb{E}[S_i|X_i] = \mathbb{E}[N_i|X_i]\mathbb{E}[Y_{i,k}|X_i]$ où X_i désigne des informations sur l'assuré i.

Une alternative est l'approche probabilité × coût. Pour l'assuré i,

$$S_i = \begin{cases} Y & \text{si } I_i = 1 \\ 0 & \text{sinon.} \end{cases}$$

où I_i est une variable aléatoire qui suit une loi de Bernoulli et vaut 1 si un sinistre survient pour l'assuré i et 0 sinon, et Y est la variable aléatoire du coût du sinistre. Les variables Y et I_i sont indépendantes. Si p_i est le paramètre de la loi de Bernoulli I_i, alors $\mathbb{E}[S_i] = p_i\mathbb{E}[Y]$.

La prime pure correspond à la prime minimale. Des chargements y sont ensuite appliqués afin d'avoir une marge de sécurité, et pour couvrir les frais associés au contrat d'assurance.

Un modèle d'estimation de la prime pure est classiquement créé pour chaque garantie, car chacune d'entre elle est supposée suivre une loi différente des autres.

3.3 Fréquence de déclaration des hôpitaux de la base contrats

La variété des incidents Cyber présume que les lois et l'effet des variables explicatives peut différer. Après sélection des incidents Cyber qui pourraient déclencher le contrat proposé par Relyens, la base sinistre contient 415 incidents entre 2014 et 2021.

Le nombre d'incidents entre 2014 et 2021 par catégorie, considérés comme un sinistre dans le contexte de Relyens, est le suivant :

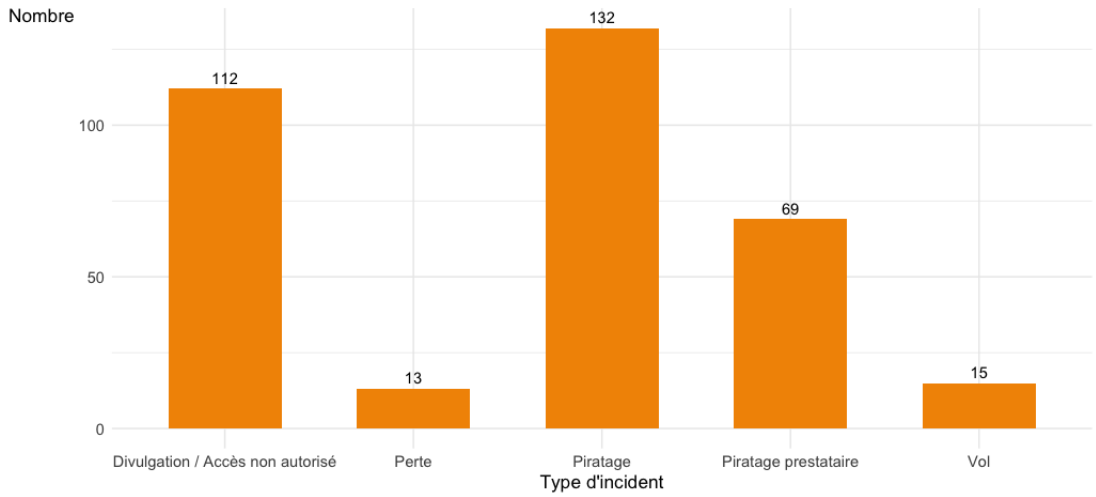


FIGURE 3.2 – Nombre de déclarations, considérées comme des sinistres, des hôpitaux par catégorie d'incident

Pour chaque catégorie, le nombre d'établissements touchés par un incident est :

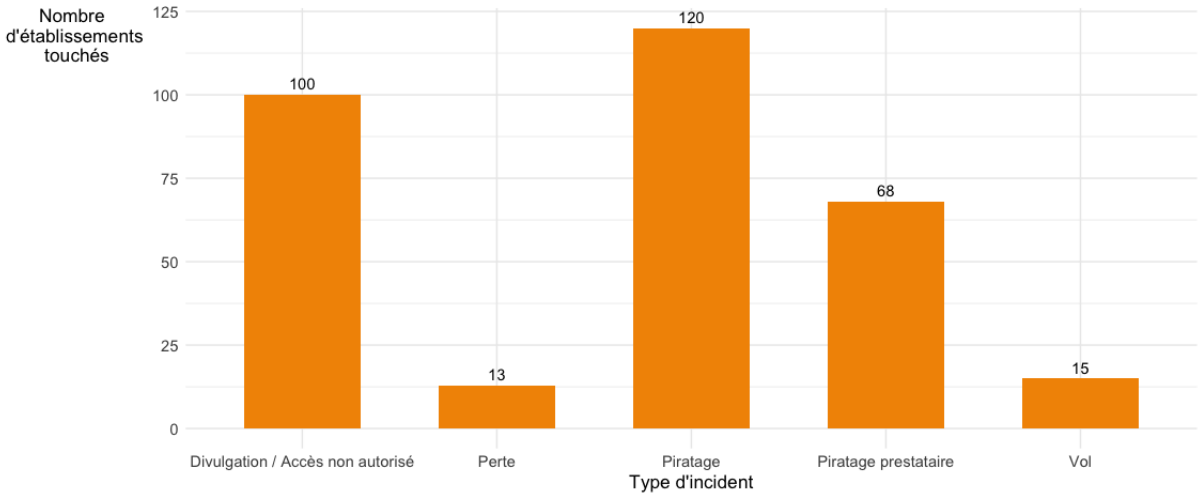


FIGURE 3.3 – Nombre d'établissements touchés au moins une fois par un certain type d'incident

Les valeurs obtenues sont très proches de celles du nombre d'incidents. Il y a en effet peu d'établissements ayant connu plusieurs incidents. Seuls 10 établissements ont déclaré plus d'un piratage au-dessus du seuil, 10 établissements une divulgation, et 1 établissement un piratage de prestataire. Les

données étant plus proches d'une variable dichotomique que d'une fréquence, il est choisi de modéliser la probabilité de survenue des incidents.

Les piratages, divulgations et piratages de prestataires sont les sinistres les plus fréquents. Peu de vols et pertes ont été déclarés, c'est pourquoi ces incidents ne feront pas l'objet d'un modèle de probabilité mais d'une estimation empirique.

Ainsi, pour ces deux incidents, la probabilité de sinistre vaut :

$$\frac{13}{6045} = 0,22\% \text{ pour les pertes}$$
$$\frac{15}{6045} = 0,25\% \text{ pour les vols}$$

Pour les incidents modélisés, la proportion d'établissements ayant eu un sinistre est aussi faible et varie entre 1% et 2%, il s'agit d'une situation de déséquilibre des classes.

3.4 Modélisation d'une variable binaire avec déséquilibre des classes

La section précédente a montré que les déclarations des hôpitaux peuvent être modélisées par une variable binaire. De plus, il y a peu d'incidents par rapport au nombre de contrats. La classe de réponses positives qui correspondent à la survenue d'un incident est donc sous-représentée, ce qui peut amener à des difficultés de calibration des modèles. Des rappels théoriques sont effectués, puis plusieurs méthodes pour faire face à ce déséquilibre sont présentées, et la plus performante est retenue.

3.4.1 Modèle linéaire généralisé pour une variable réponse binaire

Le modèle linéaire généralisé vise à expliquer une variable aléatoire Y en fonction de variables explicatives. Cette estimation est réalisée à partir des vecteurs observés des variables.

Il est considéré qu'il existe une relation de la forme suivante entre la variable à expliquer Y, et les p variables explicatives $x_1, \ldots, x_p$, :

$$g_n\left(E\left(Y|x_1, ...x_p\right)\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

avec g la fonction de lien réelle, déterministe, strictement monotone et dérivable.

Y appartient à la famille exponentielle dont la densité est :

$$f_{\theta,\phi}\left(y\right) = \exp\left(\frac{y \times \theta - b\left(\theta\right)}{\phi} + c\left(y, \phi\right)\right)$$

avec b définie sur $\mathbb{R}$ deux fois dérivable et dont la dérivée première est injective, c définie sur $\mathbb{R}^2$. θ est appelé le paramètre naturel et ϕ le paramètre de dispersion.

Lorsque la variable à expliquer Y est binaire, les fonctions de lien possibles sont les suivantes :

— Fonction de lien logit :

$$g\left(\pi\right) = \ln\left(\frac{\pi}{1-\pi}\right)$$

— Fonction de lien probit :

$$g(\pi) = \Phi^{-1}(\pi)$$

avec Φ^{-1} l'inverse de la fonction de répartition de la loi normale centrée réduite.

— Fonction de lien cloglog :

$$g(\pi) = \ln(-\ln(1-\pi))$$

L'estimation des paramètres est réalisée par maximum de vraisemblance. Pour un vecteur $Y = (y_1, \ldots, y_n)$, la fonction de log-vraisemblance est :

$$\begin{aligned} l(\theta(\beta), y, \phi) &= \sum_{i=1}^{n} \ln(f(y_i, \phi, \theta_i)) \\ &= \sum_{i=1}^{n} \frac{w_i(y_i\theta_i - b(\theta_i))}{\phi} + \sum_{i=1}^{n} c_i(y_i, \phi) \end{aligned}$$

L'estimation par maximum de vraisemblance se fait en résolvant les équations pour $j = 0, \ldots, p$:

$$\frac{\partial l(\theta(\beta), y, \phi)}{\partial \beta_j} = 0$$

3.4.2 Régression logistique et rapport des cotes

Soit P la probabilité d'un évènement étudié. Le rapport des cotes (*odds ratio*) entre deux individus x et $\hat{x}$ est défini par :

$$OR(x, \hat{x}) = \frac{\frac{P(x)}{1-P(x)}}{\frac{P(\hat{x})}{1-P(\hat{x})}}$$

L'interprétation est la suivante :

Valeur du rapport des cotes	Interprétation
$OR(x, \hat{x}) > 1$	$P(x) > P(\hat{x})$
$OR(x, \hat{x}) = 1$	$P(x) = P(\hat{x})$
$OR(x, \hat{x}) < 1$	$P(x) < P(\hat{x})$

TABLE 3.1 – Interprétation du rapport des cotes

Dans le cas du modèle de régression logistique estimant la probabilité P à partir des variables explicatives dont les valeurs observées pour un individu x sont $x_1, \ldots, x_p$, il peut être écrit la relation suivante :

$$ln\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 \times x_1 + \ldots + \beta_p \times x_p$$

avec $\beta_0, \ldots, \beta_p$ les coefficients estimés.

Donc

$$OR(x, \hat{x}) = \exp(\beta_1(x_1 - \hat{x}_1) + \ldots + \beta_p(x_p - \hat{x}_p))$$

Ainsi, si les individus x et $\hat{x}$ diffèrent seulement par la variable j,

$$OR(x, \hat{x}) = \exp(\beta_j(x_j - \hat{x}_j))$$

Le coefficient $\exp \beta_j$ mesure l'influence de la j^e variable sur le rapport des cotes. Il correspond au rapport des cotes dans le cas de l'augmentation d'une unité de la variable j.

3.4.3 Méthodes de gestion du déséquilibre des classes

La section 3.3 présente la proportion d'établissements touchés par chaque incident qui varie entre 1% et 2% pour les catégories modélisées ce qui est faible. Cela peut conduire à des difficultés de modélisation, un biais des estimateurs et une perte de précision importante dans les prédictions car cela augmente la variance.

Lors de la réalisation de prédictions afin de calculer certains indicateurs, une solution est la modification du seuil à partir duquel il y a affectation d'une réponse positive pour améliorer les prédictions. Cela signifie qu'il sera considéré qu'un incident a lieu lorsque la probabilité estimée par le modèle dépasse un seuil inférieur à 0,5. Dans ce mémoire, le seuil sera défini pour chaque modèle de façon à optimiser le F1 score, défini dans la section suivante 3.4.4.

Le mémoire d'Ornellia Djoffon [40] analyse les conséquences du déséquilibre des classes et présente des méthodes pour y remédier dans le cadre d'un modèle de régression logistique et des modèles d'apprentissage. Ce mémoire répond ainsi à des problématiques de modélisation d'une variable binaire déséquilibrée, pouvant aussi être rencontrées dans le cadre de la modélisation du risque Cyber.

Le rééchantillonnage est une méthode fréquemment employée qui a pour but de modifier le taux de la réponse positive dans les données. Différentes approches peuvent être appliquées :

— Une certaine proportion de lignes associées à une réponse négative peut être supprimée de manière aléatoire, il s'agit de l'*undersampling* ou sous-échantillonnage.

— L'*oversampling* ou sur-échantillonnage désigne la duplication aléatoire de lignes associées à une réponse positive.

— Il est également possible de combiner les deux approches précédentes, c'est le *both sampling*.

— Des données artificielles peuvent être créées. ROSE (*Random Oversampling Examples*) génère des données artificielles pour les variables quantitatives à partir de l'estimation par noyau de la densité conditionnelle des prédicteurs numériques par rapport à la réponse, et conserve les modalités prises par les variables qualitatives. Une autre méthode est SMOTE (*Synthetic Minority Oversampling Technique*) qui génère des données synthétiques par la méthode des K plus proches voisins. Cette méthode nécessite sous R un retraitement des variables qualitatives en facteurs.

Il sera vu qu'il est préférable dans le cadre de ce mémoire de discrétiser les variables, les méthodes ROSE et SMOTE ne sont donc pas retenues. Les approches *undersampling*, *oversampling*, et *both sampling* seront utilisées, puis comparées sur des critères de qualité de prédiction, et le meilleur modèle sera retenu pour chaque type d'incident.

En modifiant le taux de réponses positives, la probabilité estimée par le modèle est elle aussi modifiée. Deux approches permettent de la corriger.

40. Ornellia Djoffon, Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel, *Institut des Actuaires*, 2017

La méthode de pondération [41] applique un poids à la vraisemblance avant de la maximiser à partir des données rééquilibrées. La vraisemblance s'écrit donc :

$$\ln\left(L_w\left(p,(y,x)\right)\right) = \sum_{i=1}^{l} w\left(y_i\right) \ln\left(f\left(y_i|x_i,p\right)\right)$$

avec y la variable réponse, x les variables explicatives, $y_1, \dots y_l$ et $x_1, \dots, x_l$ les observations associées, p la probabilité modélisée, $w\left(y_i\right)$ le poids de l'observation y_i.

Les poids valent $w\left(1\right) = \frac{\tau}{\tau^c}$ et $w\left(0\right) = \frac{1-\tau}{1-\tau^c}$, où τ désigne la proportion de 1 dans les données et τ^c la proportion de 1 dans les donnés rééquilibrées.

La seconde méthode de correction, l'ajustement préalable [42], n'est applicable que dans le cas d'une régression logistique car elle provient de l'expression de la loi logit. Il s'agit de corriger le coefficient β_0 après estimation du modèle par

$$\beta_0 = \hat{\beta}_0 - \ln\left(\frac{1-\tau}{\tau}\frac{\tau^c}{1-\tau^c}\right)$$

3.4.4 Critères de choix du meilleur modèle

Les sections précédentes illustrent que plusieurs fonctions de lien peuvent être utilisées pour modéliser une variable binaire, et différentes techniques de rééchantillonnage peuvent être utilisées pour faire face au manque de données. Ces choix peuvent avoir une influence sur la qualité d'ajustement des modèles. Différents critères existent pour mesurer la qualité d'un modèle et sont calculables directement à partir de celui-ci, par exemple :

— Le critère d'information d'Akaike (AIC) mesure la qualité du modèle à partir de la vraisemblance $L\left(\theta\right)$ pénalisée par le nombre de paramètres k. Il se calcule par

$$AIC = -2\ln\left(L\left(\theta\right)\right) + 2k$$

— Le critère d'information bayésien (BIC) mesure également la qualité du modèle à partir de la vraisemblance, et la pénalise par le nombre de paramètres k et la taille de l'échantillon N. Il vaut ainsi

$$BIC = -2\ln\left(L\left(\theta\right)\right) + k\ln\left(N\right)$$

— Le R2 de McFadden est compris entre 0 et 1 et mesure la variance expliquée par le modèle. En notant L la vraisemblance estimée, M le modèle complet et M_0 le modèle sans les prédicteurs,

$$R^2_{MF} = 1 - \frac{\ln\left(L\left(M\right)\right)}{\ln\left(L\left(M_0\right)\right)}$$

Il est généralement recommandé de diviser les données en une base d'apprentissage servant à la construction du modèle, et en une base test à partir de laquelle sont réalisées les prédictions. Différents indicateurs peuvent alors être calculés pour mesurer la qualité d'ajustement ou la performance du modèle à partir de la base test. Cependant, lorsque peu de données de sinistres sont disponibles, cette division peut mener à une perte d'information importante. Ainsi, chaque modèle sera ici estimé à partir de toutes les données. Les prédictions sont réalisées sur la base contrats avant rééchantillonnage, et seront utilisées pour calculer les indicateurs F1-score et AUC, introduits ci-après.

Le F1-score provient de la matrice de confusion, qui résume les prédictions effectuées dans un tableau de la forme suivante :

41. Manski, C. F., et Lerman, S. R. (1977). The Estimation of Choice Probabilities from Choice Based Samples. Econometrica, 45, 1977-1988.

42. Xie, Y. et Manski, F. (1989), The logit model and response-based samples, Sociol. Methods Res., 17(3) :283–302.

		Valeurs prédites	
		0	1
Observations	0	VN	FP
	1	FN	VP

FIGURE 3.4 – Matrice de confusion

Avec VN = vrais négatifs, ce qui désigne les prédictions négatives dont les observations sont négatives, FP = faux positifs, qui correspond aux prédictions des réponses positives alors que les observations sont négatives, FN = faux négatifs, qui regroupe les prédictions de réponses négatives alors qu'elles sont positives dans les données, et enfin VP = vrais positifs, qui sont les bonnes prédictions des valeurs positives.

Différents indicateurs peuvent être calculés à partir de cette matrice, dont les suivants.

— Taux de bonnes prédictions = $\frac{VP+VN}{VP+VN+FP+FN}$
— Sensibilité ou rappel = $\frac{VP}{VP+FN}$
— Spécificité = $\frac{VN}{VN+FP}$
— Précision = $\frac{VP}{VP+FP}$
— Taux de faux positifs = $\frac{FP}{FP+VN}$

Le F1 score se calcule de la manière suivante :

$$\text{F1-score} = \frac{VP}{VP + \frac{1}{2}(FN + FP)}$$

Il permet de prendre en compte à la fois la précision et le rappel, et est donc une mesure adaptée dans le cas d'un faible nombre d'évènements positifs.

Ces indicateurs dépendent d'un seuil à partir duquel il est considéré que la probabilité prédite est associée à la réalisation ou non de l'évènement. Généralement de 0,5, il sera optimisé dans ce mémoire afin de maximiser le F1 score.

Un second indicateur est l'aire sous la courbe ROC ou AUC (*Area Under the Curve*). La courbe ROC (*Receiver Operating Characteristic*) désigne le taux de vrais positifs ou rappel en fonction du taux de faux positifs. Sa valeur est comprise entre 0,5 et 1. L'AUC représente la discrimination qui est faite par le modèle entre les individus positifs et négatifs.

Plus l'aire est élevée, plus le pouvoir prédictif du modèle est important. De manière générale, la qualité du modèle est évaluée de la manière suivante selon la valeur de l'AUC :

AUC	Pouvoir de discrimination
$AUC = 0,5$	Pas de discrimination
$0,7 < AUC < 0,8$	Discrimination acceptable
$0,8 < AUC < 0,9$	Discrimination excellente
$> 0,9$	Discrimination exceptionnelle

TABLE 3.2 – Pouvoir prédictif selon l'AUC

Ainsi, tous les indicateurs seront utilisés pour comparer les régressions logistiques avec différentes fonctions de lien. Le F1 Score et l'AUC serviront pour comparer les prédictions issues des modèles avec application des différentes techniques de rééchantillonnage. De plus, le test d'Hosmer-Lemeshow sera appliqué aux modèles retenus, son fonctionnement est indiqué en annexe D.

Cependant, calculer les indicateurs à partir des prédictions effectuées sur la base d'apprentissage peut mener à un surapprentissage, et la fiabilité des valeurs obtenues est limitée, notamment lorsque la complexité des modèles est différente. Afin de limiter cet effet, il est possible de construire le modèle sur la totalité des données, puis d'évaluer les indicateurs par validation croisée. Pour une valeur K, la base de données est divisée aléatoirement en K sous-bases de mêmes dimensions. Le modèle est estimé sur $K-1$ sous-bases et les prédictions sont effectuées sur la K^{eme} base. Cette méthode est appliquée à toutes les combinaisons possibles, et l'indicateur estimé est la moyenne des indicateurs obtenus à chaque étape.

3.5 Préparation des données à la modélisation

3.5.1 Sélection des variables non corrélées

Afin de pouvoir ajuster un modèle GLM, il est nécessaire de sélectionner les variables au préalable afin d'éviter les problèmes de multicolinéarité. Lorsque deux variables sont fortement corrélées linéairement, il est fort probable que cela engendre une multicolinéarité.

Le graphique suivant indique le niveau de corrélation entre les variables quantitatives de la base contrats. Les couleurs bleues indiquent une corrélation positive et les couleurs rouge une corrélation négative. Plus la couleur est foncée, plus la corrélation est importante.

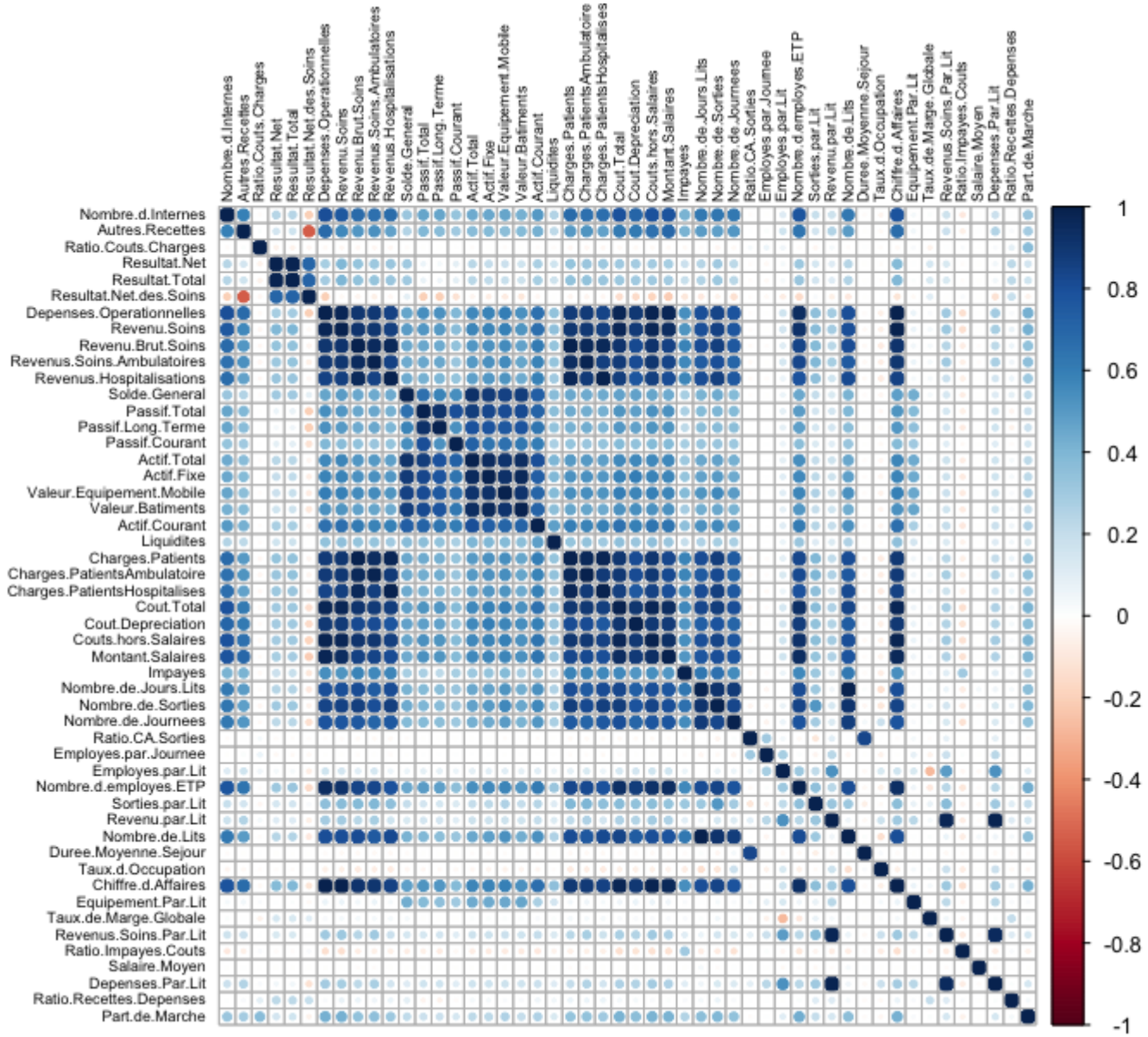


FIGURE 3.5 – Graphique des corrélations au sein de la base contrats

De fortes corrélations se trouvent dans la base contrats. En effet, le nombre de lits dépend directement du nombre d'employés et de sorties. Seules des variables dont la corrélation n'excède pas 60% sont conservées. Ce seuil a été choisi afin de ne pas avoir de nombreuses variables fortement ou très fortement corrélées entre elles dans le modèle. Le choix des variables à conserver dépend du taux de données manquantes et de l'objectif final du mémoire qui est une tarification pour des établissements en France. La variable doit donc être dans l'idéal comparable à une donnée disponible en France. Cependant, les variables non comparables sont tout de même conservées dans un premier temps afin de mesurer leurs effets.

Le test d'indépendance du χ^2 est appliqué entre les variables qualitatives. Son fonctionnement est rappelé en annexe D. Il indique que toutes les variables qualitatives sont liées car la p valeur de chaque test est nulle.

Le coefficient de Cramer est basé sur le χ^2 et permet de quantifier le lien entre deux variables

qualitatives. Il est défini par

$$\sqrt{\frac{\chi^2}{n \times min(q_1 - 1, q_2 - 1)}}$$

avec n le nombre de données, q_1 le nombre de modalités de la première variable, et q_2 celui de la seconde variable.

La variable Type d'établissement selon la classification de *CMS* est retirée du modèle car elle est très liée avec plusieurs autres variables, et ses modalités sont très similaires à celles de la variable Type de soins.

D'autres liens forts sont mis en avant entre plusieurs variables, notamment entre les couples de variables suivants : Hôpital accès critique et Rural ou Urbain, Type de soins et Certification données, Service d'urgences et Certification données, Type de soins et Service d'urgences. Les variables sont tout de même conservées dans un premier temps afin de déterminer par la suite les plus influentes dans les modèles.

	1	2	3	4	5	6	7	8	9	10
1 - Rural ou Urbain	1,00	0,26	0,33	0,60	0,36	0,35	0,11	0,33	0,25	0,42
2 - Accueille des internes	0,26	1,00	0,14	0,28	0,21	0,26	0,32	0,14	0,15	0,29
3 - Logiciel externe données	0,33	0,14	1,00	0,26	0,42	0,89	0,41	0,71	0,75	0,26
4 - Hopital Acces Critique	0,60	0,28	0,26	1,00	0,34	0,34	0,17	0,28	0,20	0,44
5 - Type de Contrôle	0,36	0,21	0,42	0,34	1,00	0,34	0,30	0,43	0,36	0,35
6 - Type de Soins	0,35	0,26	0,89	0,34	0,34	1,00	0,30	0,77	0,73	0,13
7 - Membre Système	0,11	0,32	0,41	0,17	0,30	0,30	1,00	0,36	0,40	0,24
8 - Service d'urgences	0,33	0,14	0,71	0,28	0,43	0,77	0,36	1,00	0,60	0,23
9 - Certification données	0,25	0,15	0,75	0,20	0,36	0,73	0,40	0,60	1,00	0,24
10 - Etat	0,42	0,29	0,26	0,44	0,35	0,13	0,24	0,23	0,24	1,00

TABLE 3.3 – Coefficients de Cramer

Le rapport de corrélation est calculé entre les variables quantitatives retenues et les variables qualitatives. Pour une variable qualitative avec k modalités, une variable quantitative X, et i observations, il s'exprime par

$$\eta^2 = \frac{\sum_k n_k(\overline{x_k} - \overline{X})^2}{\sum_i (x_i - \overline{X})^2}$$

où n_k est le nombre d'individus de la modalité k, $\bar{x_k}$ est la moyenne de X pour ce sous-groupe d'individus, $\bar{X}$ est la moyenne de X.

Ce rapport met en avant qu'il n'y a pas de liens forts entre les variables quantitatives et les variables qualitatives de la base, comme le montre le tableau ci-dessous.

	1	2	3	4	6	7	8	9	10
Nombre d'employés ETP	0,03	0,26	0,03	0,06	0,07	0,11	0,04	0,03	0,05
Impayés	0,03	0,14	0,03	0,05	0,04	0,09	0,03	0,03	0,07
Liquidités	0,00	0,02	0,00	0,00	0,02	0,01	0,00	0,00	0,01
Actif Total	0,01	0,07	0,01	0,02	0,04	0,04	0,01	0,01	0,05
Résultat Net	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,03
Ratio Coûts Charges	0,00	0,00	0,01	0,00	0,01	0,01	0,00	0,01	0,73
Taux d'Occupation	0,08	0,01	0,01	0,14	0,02	0,00	0,01	0,01	0,07
Ratio Recettes Dépenses	0,00	0,00	0,00	0,00	0,02	0,00	0,01	0,00	0,01
Sorties par Lit	0,06	0,13	0,09	0,13	0,14	0,18	0,06	0,10	0,09
Durée Moyenne Séjour	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01
Dépenses Par Lit	0,00	0,03	0,02	0,00	0,19	0,02	0,03	0,02	0,06
Ratio Salaires Coûts	0,00	0,00	0,01	0,00	0,04	0,01	0,01	0,01	0,01
Ratio Impayés Coûts	0,00	0,01	0,00	0,01	0,02	0,00	0,00	0,00	0,21
Part de Marche	0,00	0,07	0,01	0,02	0,02	0,03	0,02	0,01	0,49
Résultat Net	0,00	0,01	0,00	0,00	0,02	0,01	0,01	0,01	0,02
Équipement Par Lit	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,00	0,02
Ratio CA Sorties	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Employés par Journée	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01
Employés par Lit	0,00	0,01	0,01	0,00	0,06	0,01	0,02	0,01	0,04
Revenu par Lit	0,00	0,03	0,02	0,00	0,19	0,02	0,03	0,02	0,06
Taux de Marge Globale	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Revenus Soins Par Lit	0,00	0,03	0,03	0,00	0,18	0,03	0,03	0,03	0,07
Salaire Moyen	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

TABLE 3.4 – Rapports de corrélation

Ainsi, les variables du modèle sont les suivantes :

— Nombre d'employés ETP
— Impayés
— Liquidités
— Actif total
— Résultat net
— Ratio Coûts Charges
— Ratio Impayés Coûts
— Part de marché
— Salaire moyen
— Équipement par lit
— Ratio CA Sorties
— Employés par journée
— Employés par lit
— Taux Marge Globale
— Etat
— Rural ou Urbain
— Accueille des internes
— Logiciel externe données
— Certification données
— Hôpital d'accès critique
— Type de soins

— Membre d'un système
— Service d'urgences
— Taux d'occupation
— Dépenses par lit
— Ratio Recettes Dépenses
— Sorties par lit

Une sélection des variables les plus influentes sera réalisée par la suite.

3.5.2 Identification de profils de risque à partir d'une ACP et ACM

Avant de commencer la modélisation, il peut être intéressant d'analyser les données avec les variables sélectionnées. L'analyse en composante principale (ACP) et l'analyse des correspondances multiples (ACM) sont appliquées afin de révéler des profils de risque à partir de ces variables.

D'abord, l'analyse en composante principale permet d'explorer les données en résumant les informations de variables quantitatives, à partir de la construction de nouvelles variables appelées composantes principales. Elle permet la projection de l'information sur un plan factoriel, chaque axe correspondant à une composante principale. L'information résumée sur le premier plan factoriel est la suivante :

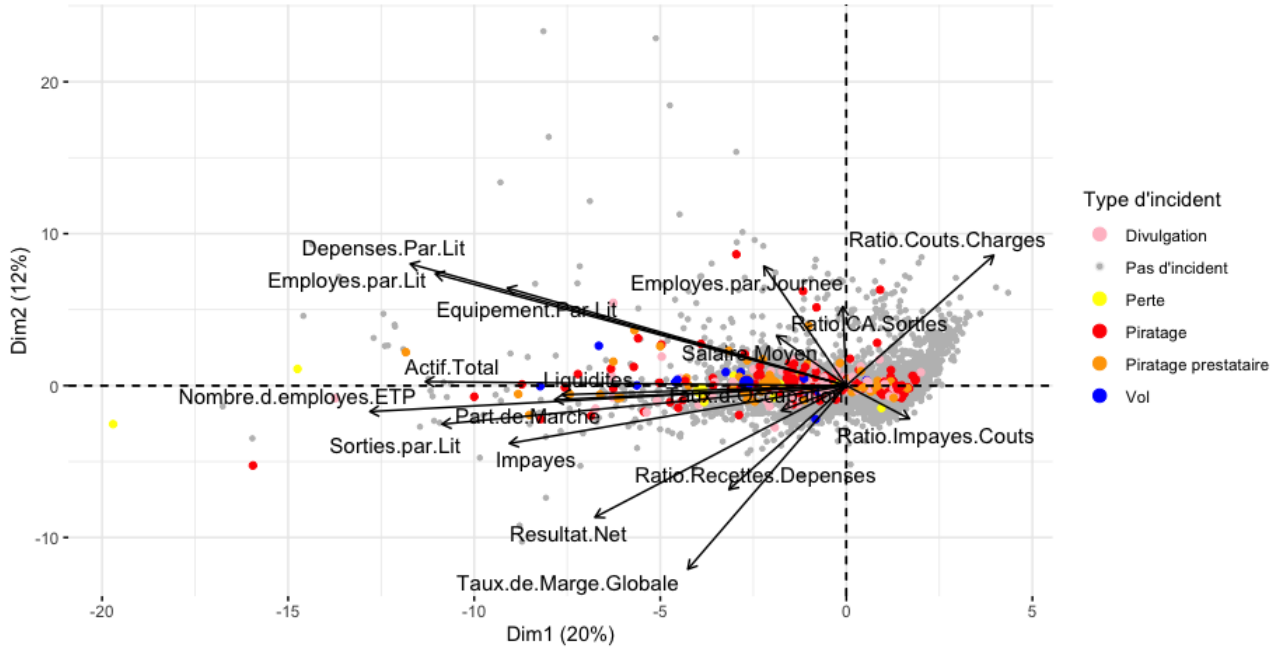


FIGURE 3.6 – Analyse en composante principale axes 1 et 2

Chaque établissement est représenté par un point, gris s'il n'a pas déclaré d'incident ou de couleur s'il en a déclaré un. Les variables sont représentées par les flèches. Les incidents semblent assez répartis, mais ils restent plutôt alignés sur l'axe horizontal, qui est bien expliqué par les variables Nombre d'employés ETP, Actif Total et Sorties par lit par exemple. Il est possible de s'attendre à ce que la probabilité d'incident soit liée plus particulièrement à ces variables. Un certain nombre d'établissements ayant eu un incident se trouvent à gauche de la représentation, et sont donc associés à un nombre élevé pour les variables Nombre d'Employés ETP, Actif Total, Sorties par Lit, Part de Marché et Liquidités.

Ensuite, les variables qualitatives sont représentées par le premier plan factoriel d'une ACM, permettant également de résumer à l'aide de variables construites les informations issues de variables qualitatives :

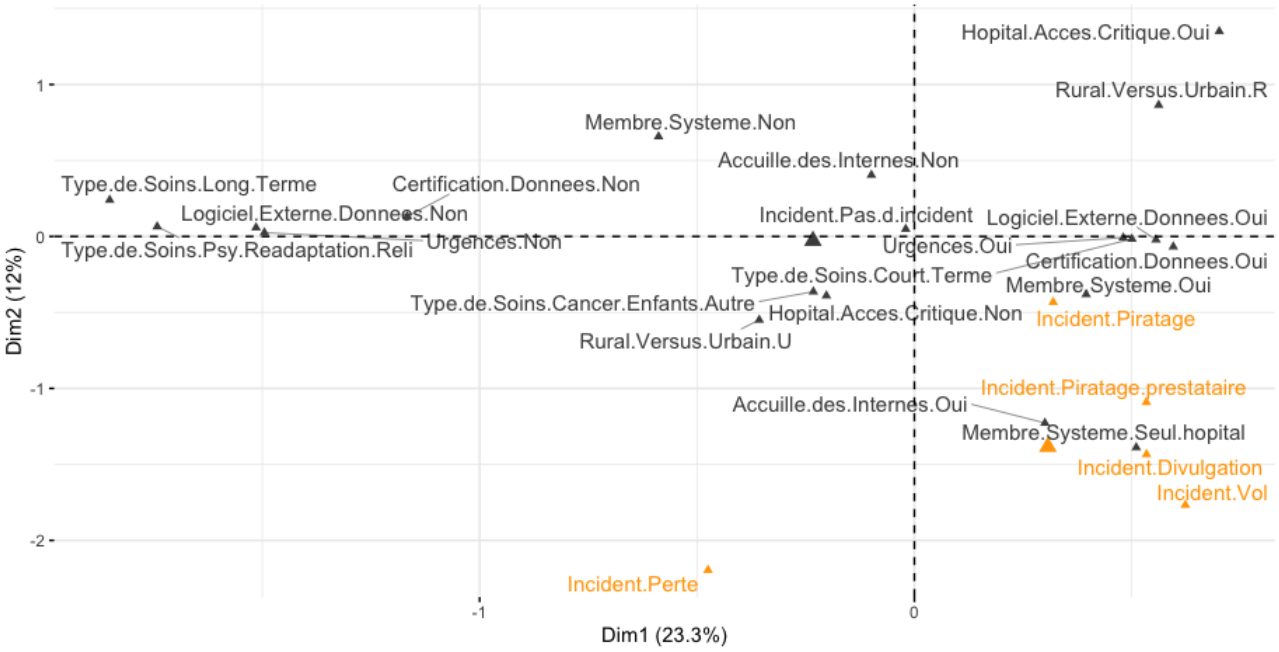


FIGURE 3.7 – Analyse des correspondances multiples axes 1 et 2

La variable Incident désigne ici le type d'incident subi par un établissement. Il est possible de voir que le vol, la divulgation, le piratage de prestataires et le piratage sont situés du même côté des axes, avec les modalités Accueille des internes, Seul hôpital membre du système, et membre du système oui (avec d'autres hôpitaux). Ces modalités semblent donc particulièrement associées à la probabilité d'incident. Au contraire, le premier axe oppose les incidents aux types de soins de long terme, de réadaptation, psychiatriques et les institutions religieuses non médicales, qui devraient donc être liés à un risque faible d'incident.

3.5.3 Regroupement de modalités et discrétisation des variables quantitatives

Différents regroupements et discrétisations sont réalisés sur la base initiale, avant application des méthodes de rééchantillonnage.

La variable type d'appartenance est regroupée pour représenter les catégories Public, Privé, et Non lucratif.

Les proportions d'hôpitaux déclarant un incident au sein de chaque Etat sont calculées pour chaque catégorie d'incident, en divisant le nombre d'hôpitaux ayant fait une déclaration pour un Etat par le nombre d'hôpitaux de la base contrats qui se trouvent dans le même Etat. Les Etats sont ainsi regroupés en trois groupes : Etats fortement touchés (proportion supérieure à 5%), Etats moyennement touchés (proportion entre 1 et 5%), Etats faiblement touchés (proportion inférieure à 1%).

Les proportions sont résumées sur les cartes suivantes :

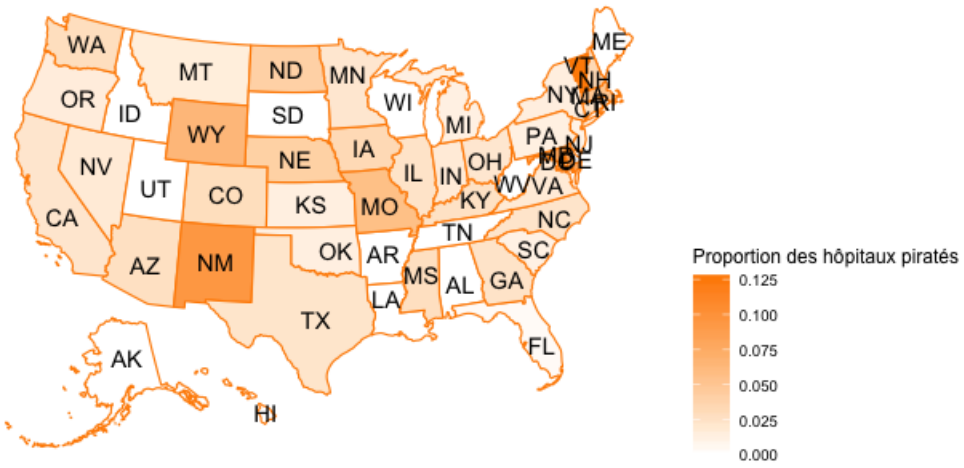


FIGURE 3.8 – Proportion des hôpitaux piratés par Etat

Les Etats à forte proportion de piratages sont dans l'ordre du plus touché au moins touché le Vermont, le New Mexico, le Maryland, le District de Columbia, le Wyoming, le Missouri, et le Massachusetts.

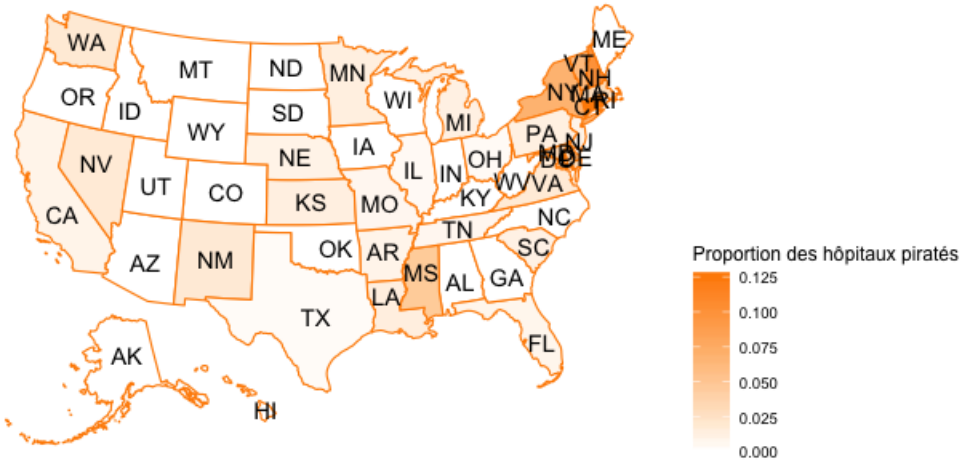


FIGURE 3.9 – Proportion des hôpitaux dont les prestataires ont été piratés par Etat

Ceux à forte proportion de piratages de prestataires sont le Connecticut, le New Hampshire, le Delaware, New York, et le Vermont, tous situés au nord est des Etats-Unis.

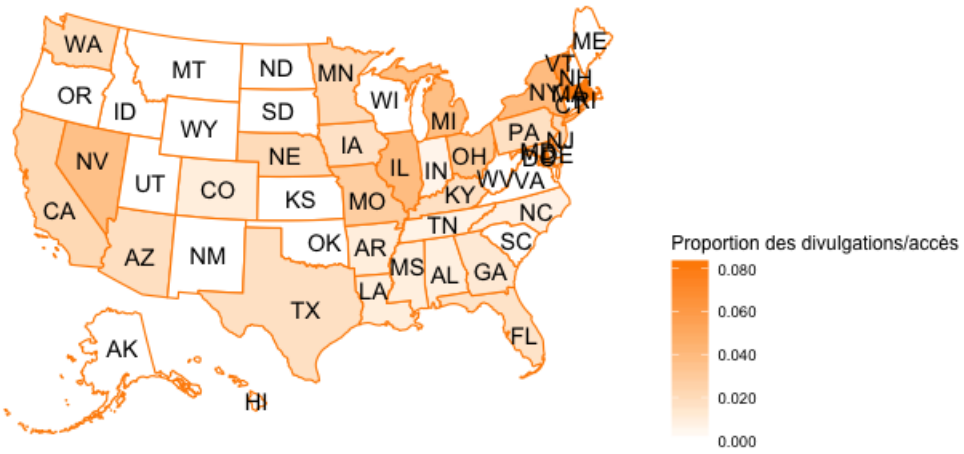


FIGURE 3.10 – Proportion des hôpitaux touchés par une divulgation ou un accès non autorisé par Etat

Concernant les divulgations, les Etats les plus touchés sont le Massachusetts, le District of Columbia, le Rhode Island, et le Vermont.

De manière générale, la probabilité de violations de données personnelles est plus élevée dans les Etats du nord-est des Etats-Unis, en particulier pour les piratages des prestataires.

Ensuite, d'autres variables catégorielles du modèle ont également fait l'objet de regroupements. Les modalités déjà présentes sont regroupées lorsqu'il est observé graphiquement que la probabilité moyenne pour les groupes associés à ces modalités sont proches.

Enfin, les variables quantitatives ont toutes été discrétisées. Il est possible d'observer lors de l'analyse exploratoire des données des effets non linéaires pour les variables quantitatives. Par exemple, le graphique suivant illustre la probabilité moyenne de piratage par classe de la variable Nombre d'employés ETP. L'ordonnée à gauche permet de mesurer le nombre de contrats dans chaque classe, et celle de droite illustre le niveau de la probabilité moyenne par classe. Les moustaches représentent l'intervalle de confiance à 95% sous l'hypothèse d'une loi normale de la valeur de la probabilité.

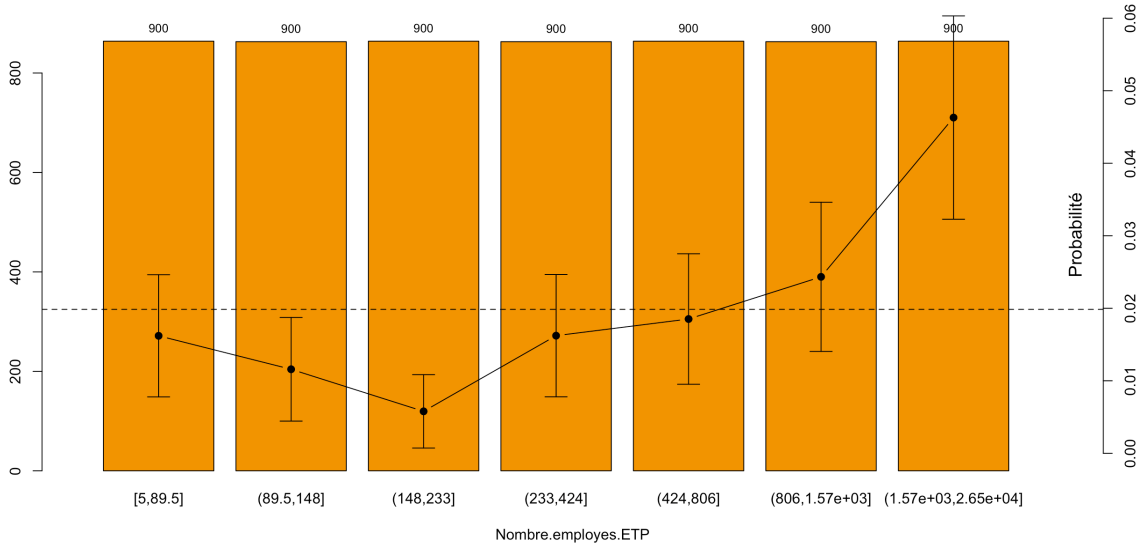


FIGURE 3.11 – Probabilité moyenne par classes de la variable Nombre d'employés ETP

Afin de prendre en compte ces effets dans le modèle, une solution est de discrétiser les variables.

Les variables quantitatives sont toutes transformées en variables catégorielles en créant des classes. Dans un premier temps, les variables sont divisées en 7 intervalles consécutifs de même effectif, et la moyenne de la variable à expliquer est calculée pour chaque intervalle, ainsi que son intervalle de confiance à 95%. Les résultats sont affichés graphiquement, comme 3.11. L'idée est de regrouper les intervalles dont la moyenne est proche. Ces graphiques sont comparés avec une deuxième méthode, où les intervalles ne sont pas de mêmes effectifs mais ont été construits par un algorithme en fonction de la variable à expliquer.

Cet algorithme est implémenté afin de réaliser une répartition optimale en se basant sur la mesure *Weight Of Evidence (WOE)* et la valeur de l'information. Couramment utilisées en risque de crédit, ces mesures permettent d'illustrer les relations entre une variable explicative et une variable réponse binaire. L'idée est de discrétiser une variable en fonction de cette relation.

Pour une discrétisation en classes $B_1, \ldots, B_n$ d'une variable explicative X par rapport à la variable réponse binaire Y, la mesure WOE associée à la modalité i correspond à

$$WOE_i = \ln\left(\frac{P(X \in B_i|Y=1)}{P(X \in B_i|Y=0)}\right)$$

La valeur de l'information de la variable X est ensuite définie par

$$VI = \sum_{i=1}^{k}(P(X \in B_i|Y=1) - P(X \in B_i|Y=0)) \times WOE_i$$

Selon sa valeur, il est possible de catégoriser le pouvoir prédictif de la variable, comme le résume le tableau ci-dessous.

Valeur de l'information	Pouvoir prédictif
$< 0{,}02$	Sans intérêt pour la prédiction
0,02 - 0,1	Faible prédicteur
0,1 - 0,3	Prédicteur moyen
$> 0{,}3$	Prédicteur fort

TABLE 3.5 – Pouvoir prédictif selon la valeur de l'information

Une approche par arbre est implémentée et les segmentations sont effectuées de manière itérative. La segmentation s'arrête lorsque la valeur de l'information apporte un gain inférieur à une valeur paramétrable. La valeur par défaut, 0,1, a été conservée.

Le graphique ci-dessous présente un exemple des classes obtenues après utilisation de cet algorithme :

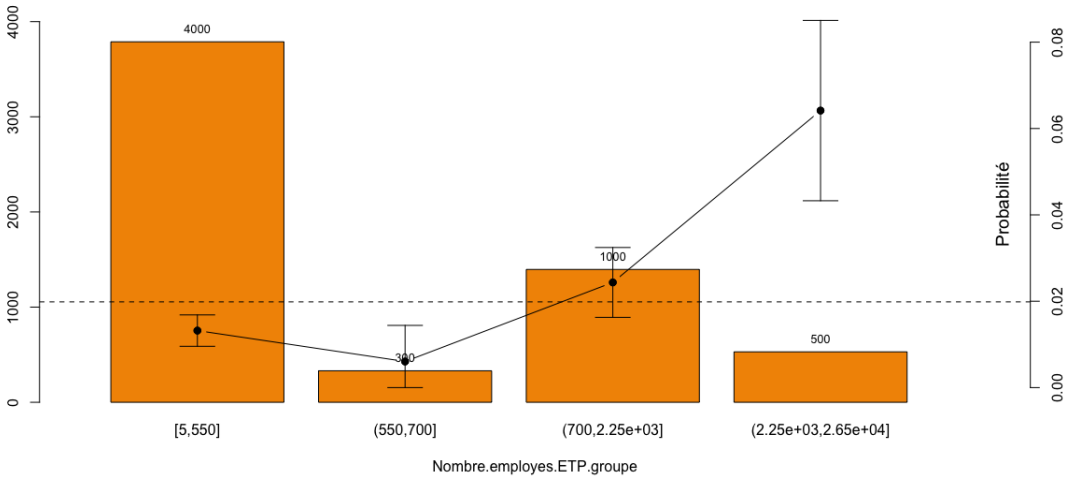


FIGURE 3.12 – Probabilité moyenne de piratage par modalité de la variable Nombre d'employés ETP

Par souci de cohérence, des regroupements supplémentaires pourront être effectués. Par exemple, il s'agit ici de regrouper en une modalité les hôpitaux ayant entre 5 et 700 employés, afin d'obtenir une croissance cohérente de la probabilité avec les effectifs des employés. Ce regroupement supplémentaire sera effectué uniquement sur le modèle final retenu dans une optique d'efficacité.

Il est également possible d'observer pour chaque variable discrétisée par la mesure *WOE* sa valeur d'information ci-dessous.

Variable (catégorielle)	Piratage	Piratages prestataires	Divulgation
Équipement par Lit	0,64	0,79	0,58
Dépenses par Lit	0,6	0,96	0,82
Liquidités	0,52	0,92	0,34
Employés par lit	0,5	1,16	0,71
Actif	0,49	0,9	0,9
Nombre d'employés ETP	0,39	1,11	1,15
Ratio CA Sorties	0,35	0,31	0,3
Part de marché	0,34	1,01	0,62
Salaire Moyen	0,33	0,59	0,19
Employés par journée	0,32	0,52	0,58
Membre d'un système	0,3	0,41	0,39
Taux marge globale	0,28	0,77	0,14
Ratio Recettes Dépenses	0,27	0,74	0,2
Type de Soins	0,26	0,53	0,54
Impayés	0,25	0,45	0,82
Etat	0,25	1,29	0,7
Ratio Coûts Charges	0,25	0,58	0,43
Ratio Impayés Coûts	0,22	0,59	0,2
Sorties par lit	0,18	0,52	0,61
Taux d'occupation	0,17	0,38	0,24
Service d'urgences	0,1	0,24	0,19
Résultat net	0,1	0,31	0,68
Logiciel externe données	0,05	0,09	0,11
Accueille des internes	0,05	0,39	0,62
Certification données	0,03	0,05	0,1
Hôpital d'accès critique	0,01	0,01	0,21
Rural ou Urbain	0,003	0,003	0,08

TABLE 3.6 – Valeurs de l'information de chaque variable pour les modèles américains

Selon cette mesure, la variable la plus influente sur la probabilité de piratage est Équipement par lit. Celle qui influe le plus la probabilité de piratage de prestataires est Etat, et Nombre d'employés ETP est l'information la plus influente sur la probabilité de divulgations.

Après ces regroupements, il n'y a plus de variables quantitatives. Il est vérifié que les variables ne sont pas trop liées entre elles, en effectuant un test du khi-deux et en observant le coefficient de Cramer. Les p values du test du khi-deux sont nulles, ce qui signifie que les variables sont liées.

Le coefficient de Cramer est supérieur à 0,7 pour les couples de variables suivants :
— Ratio Recettes Dépenses et Taux de marge globale
— Type de soins, Logiciel externe de données, Certification de données et Service d'urgences

Il est choisi de retirer certaines des variables ci-dessus pour ne pas avoir trop de liens entre elles. Les variables conservées sont celles qui semblent avoir une influence plus importante sur la probabilité en observant les graphiques des probabilités moyennes par modalité. Par exemple, en comparant les variables Type de soins et Service d'urgences, avec les graphiques ci-dessous, il est possible d'observer que les probabilités moyennes de la variable Type de soins sont plus différenciées en fonction des modalités que celles de la variable Service d'urgences.

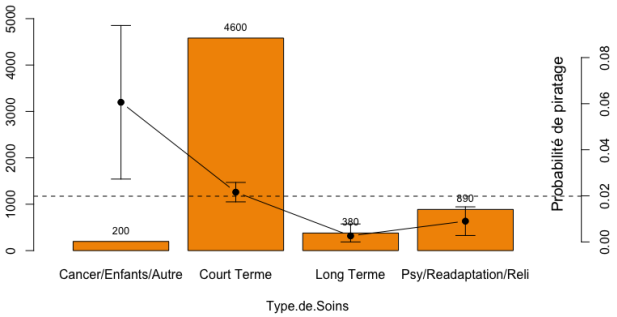


FIGURE 3.13 – Probabilité de piratage selon le type de soins

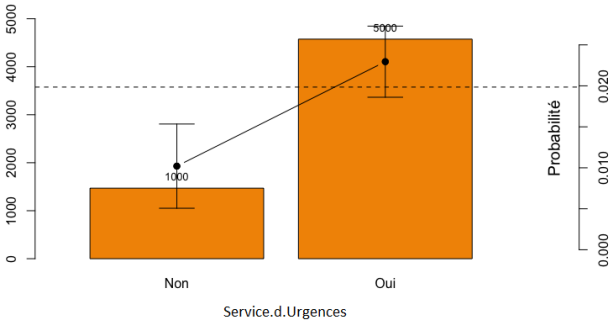


FIGURE 3.14 – Probabilité de piratage selon la variable Service d'urgences

Utiliser un logiciel d'un fournisseur pour gérer les données, être certifié pour l'utilisation des données électroniques et avoir un service d'urgences sont des modalités qui augmentent la probabilité de piratage, de piratage de prestataires et de divulgations. Cependant, cette hausse est faible et ne permettra pas à la variable d'être significative dans le modèle. Par exemple, les graphiques suivants présentent les probabilités moyennes pour chaque incident en fonction de la variable Certification données.

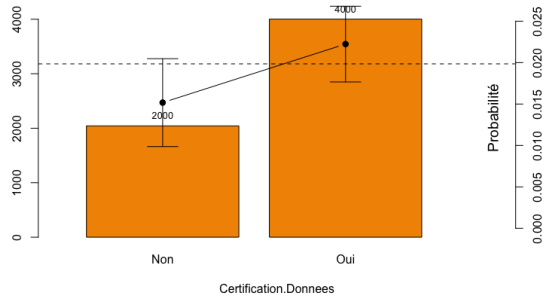


FIGURE 3.15 – Probabilité de piratage Certification données

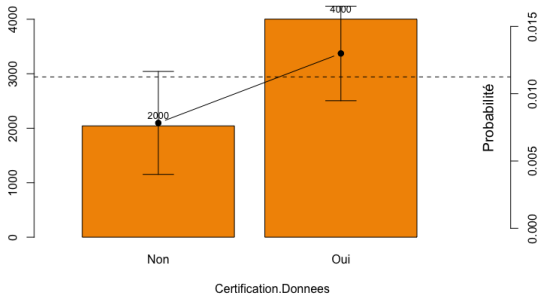


FIGURE 3.16 – Probabilité de piratage de prestataire Certification données

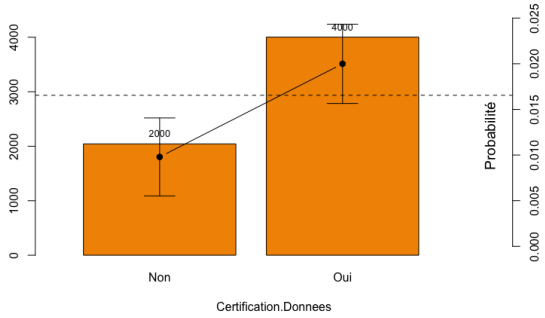


FIGURE 3.17 – Probabilité de divulgation Certification données

Afin de pouvoir apprécier l'amélioration de la qualité des modèles grâce à la discrétisation des variables quantitatives, deux modèles logistiques sont estimés : un à partir des variables continues et qualitatives, le second à partir des variables toutes discrétisées. Le graphique suivant présente la courbe ROC des deux modèles.

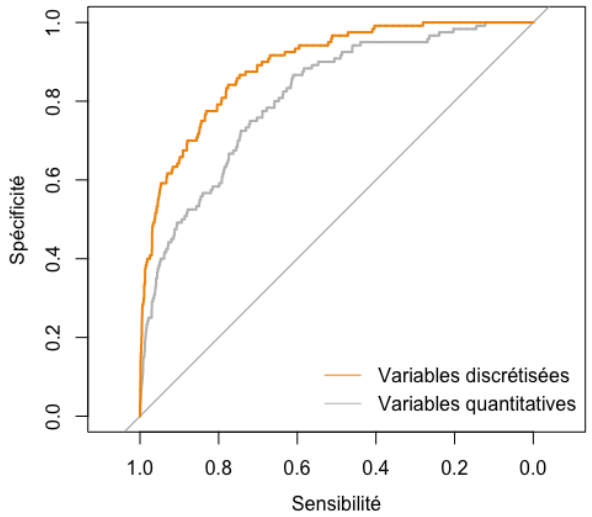


FIGURE 3.18 – Courbe ROC piratages

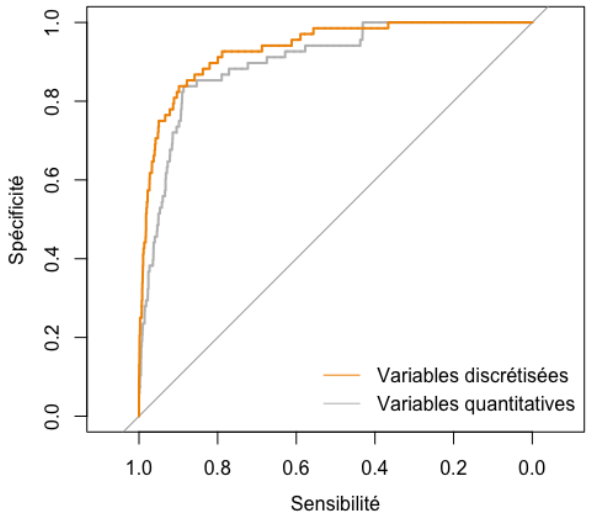


FIGURE 3.19 – Courbe ROC piratages prestataires

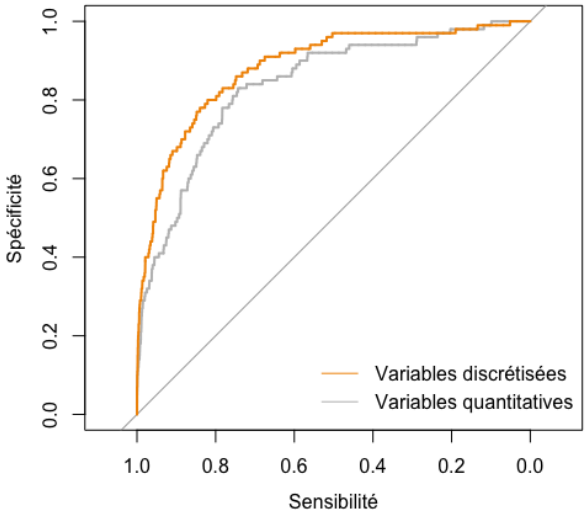


FIGURE 3.20 – Courbe ROC divulgations

Il est possible d'observer que l'aire sous la courbe du modèle avec variables discrétisées est plus grande que celle du modèle avec les variables continues. La qualité du modèle est donc meilleure grâce à la discrétisation.

3.5.4 Choix de la fonction de lien

La fonction de lien logistique est généralement préférée car appliquer la fonction exponentielle aux coefficients permet d'obtenir directement les rapports des cotes. Elle permet une interprétation quasi directe des résultats et sera donc utilisée dans ce mémoire. Afin de vérifier que les autres fonctions de lien possibles, à savoir probit et loglog, n'améliorent pas de manière très importante la qualité du modèle, un modèle de régression logistique est estimé à partir de toutes les données discrétisées pour chaque fonction de lien. Les indicateurs permettant la comparaison des modèles sont ensuite calculés et résumés ci-dessous :

Fonction de lien	R2 de McFadden	F1-Score	AUC	AIC	BIC
Logit	0,2855	0,3644	0,891	997,8	1521
Probit	0,28149	0,3346	0,895	1003	1526
Loglog	0,28623	0,3675	0,889	997	1520

TABLE 3.7 – Régression logistique probabilité de piratage pour les hôpitaux américains

Fonction de lien	R2 de McFadden	F1-Score	AUC	AIC	BIC
Logit	0,36725	0,3522	0,934	605,7	1055
Probit	0,3613	0,3399	0,94	610,1	1059
Loglog	0,36973	0,3452	0,932	603,9	1053

TABLE 3.8 – Régression logistique probabilité de piratage de prestataires pour les hôpitaux américains

Fonction de lien	R2 de McFadden	F1-Score	AUC	AIC	BIC
Logit	0,26695	0,3295	0,883	890,8	1374
Probit	0,25663	0,2896	0,886	901,3	1384
Loglog	0,26899	0,3212	0,882	888,7	1372

TABLE 3.9 – Régression logistique probabilité de divulgation/accès non autorisé pour les hôpitaux américains

Les valeurs des indicateurs pour les différentes fonctions de lien sont très proches, aucune fonction de lien ne semble donc améliorer les modèles de manière significative par rapport aux autres. La fonction logit présente l'avantage de permettre une interprétation facile des coefficients avec les rapports des cotes, et est donc retenue pour la modélisation.

3.5.5 Sélection des variables influentes

Afin de mesurer l'influence globale des variables catégorielles, une analyse de la variance est effectuée sur le logiciel R. Cette analyse compare un modèle restreint H_0 et un modèle complet H_1. Elle se base sur le test du rapport de vraisemblance. La statistique du test est

$$\Delta = D_0 - D_1 = 2\left(l_{\hat{\beta}_1} - l_{\hat{\beta}_0}\right) \sim \chi^2_q$$

où q est la différence du nombre de paramètres du modèle complet et du modèle restreint, D_0 la déviance du modèle restreint, D_1 la déviance du modèle complet, $l_{\hat{\beta}_1}$ la logvraisemblance du modèle complet et $l_{\hat{\beta}_0}$ la logvraisemblance du modèle restreint.

Trois types d'analyses de variance existent :

— L'analyse de type I teste les variables les unes après les autres. Chaque variable est ajoutée au modèle restreint, sous l'hypothèse que les variables explicatives précédentes font partie du modèle restreint. Cet ajout forme le modèle complet.

— L'analyse de type II teste chaque variable explicative quand toutes les autres variables sont présentes dans le modèle restreint. Elle respecte le principe de marginalité qui indique qu'en cas de variables croisées présentes dans le modèle testé, les variables marginales doivent également être incluses dans le modèle.

— Le type III teste également chaque variable explicative avec toutes les autres présentes dans le modèle, sans respecter le principe de marginalité. La différence entre ces deux tests est donc la gestion des termes d'interaction.

La documentation R recommande l'utilisation de l'analyse de type II, ce qui correspond à l'analyse de type II menée par SAS lorsque les variables sont catégorielles. Cette analyse sera donc utilisée pour la sélection des variables à inclure dans le modèle. Une analyse de type III a également été réalisée et retenait les mêmes variables influentes.

La significativité des coefficients de chaque modalité est ensuite étudiée, et des modalités proches peuvent être regroupées. Le test de significativité est le test de Wald.

L'hypothèse testée est $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$ dont la statistique est la suivante :

$$W = \frac{\hat{\beta}_j^2}{var\left(\hat{\beta}_j\right)} \sim \chi^2(1)$$

La sélection est d'abord réalisée à partir du modèle logistique classique. Les variables influentes retenues pour chaque modèle sont résumées dans les tableaux ci-dessous :

Variable	LR Chisq	Df	Pr(>Chisq)	Valeur de l'information
Etat	50,24	2	0	0,25
Membre d'un système	23,73	2	0	0,3
Employés par journée	26,42	5	0	0,32
Actif total	19,77	3	0,0002	0,49
Taux de marge globale	21,89	5	0	0,28
Ratio Coûts Charges	19,83	4	0	0,25
Taux d'occupation	22,39	6	0	0,17
Accueille des internes	6,19	1	0,01	0,05
Salaire moyen	17,44	6	0,01	0,33
Ratio CA Sorties	10,09	3	0,02	0,35
Impayés	9,86	3	0,02	0,25
Part de marché	9,47	3	0,03	0,34
Sorties par lit	10,18	4	0,04	0,18

TABLE 3.10 – Variables influentes piratages

Variable	LR Chisq	Df	Pr(>Chisq)	Valeur de l'information
Etat	56,35	2	0,0000	1,29
Taux de Marge globale	16,93	3	0,0007	0,77
Employes par lit	12,83	3	0,0050	1,16
Ratio Coûts Charges	14,76	4	0,0052	0,58
Ratio Impayés Coûts	8,63	3	0,0347	0,59
Impayés	8,75	3	0,0328	0,45
Membre d'un système	7,64	2	0,0219	0,41
Employés par journée	11,86	4	0,0185	0,52
Liquidités	6,93	3	0,0743	0,92

TABLE 3.11 – Variables influentes piratage prestataire

Variable	LR Chisq	Df	Pr(>Chisq)	Valeur de l'information
Etat	46,46	2	0,0000	0,7
Employes par Lit	22,81	3	0,0000	0,71
Taux d'Occupation	21,22	6	0,0017	0,24
Impayés	15,59	4	0,0036	0,82
Ratio CA Sorties	11,33	3	0,0101	0,3
Ratio Coûts Charges	14,23	6	0,0271	0,43
Type de Soins	8,57	3	0,0357	0,54
Hopital d'Accès Critique	3,89	1	0,0485	0,21
Accueille des internes	3,53	1	0,0601	0,62

TABLE 3.12 – Variables influentes divulgations

Les valeurs de l'information sont toutes supérieures à 0,1, sauf la variable Accueille des internes pour les piratages qui vaut 0,05 et est donc considéré comme un faible prédicteur.

Plusieurs modèles sont ensuite estimés à partir de bases d'apprentissage formées par les méthodes de rééchantillonnage présentées dans la section 3.4.3, à savoir le sous-échantillonnage, le suréchantillonnage, et la combinaison de sous et suréchantillonnage. Il est choisi d'aboutir à une proportion de 50% de réponses positives pour le suréchantillonnage, et la combinaison de sous et suréchantillonnage, afin d'avoir une proportion équilibrée. Une proportion de 30% est retenue pour le sous-échantillonnage dans le but de ne pas supprimer un trop grand nombre d'observations avec une réponse nulle. Les deux méthodes de correction du rééchantillonnage également mentionnées précédemment sont utilisées : la méthode de correction par poids et l'ajustement préalable.

Pour certaines méthodes, des variables supplémentaires peuvent apparaître comme influentes dans le modèle. Les modèles sont d'abord estimés à partir des variables indiquées ci-dessus, influentes dans les données sans rééchantillonnage. Puis dans un second temps, une nouvelle sélection de variables influentes est réalisée pour chaque modèle estimé à partir des bases rééchantillonnées, et les modèles sont réestimés. Pour chaque modèle, un seuil d'affectation à une réponse positive est déterminé de manière à maximiser le F1-Score. Les modèles sont comparés entre eux par le F1-Score obtenu et l'AUC. Le F1-Score est l'indicateur qui est regardé en priorité, car il se concentre sur les prédictions des réponses positives, alors que l'AUC concerne à la fois les prédictions des réponses positives et négatives.

Données	Correction	Piratage			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,33	0,15	0,88	0,30	0,19	0,92	0,27	0,14	0,87
Over	Poids	0,32	0,16	0,88	0,30	0,16	0,92	0,27	0,14	0,87
Under	Poids	0,25	0,35	0,85	0,20	0,49	0,89	0,19	0,31	0,85
Both	Poids	0,35	0,19	0,87	0,26	0,18	0,92	0,27	0,16	0,87
Over	Préalable	0,27	0,19	0,88	0,24	0,12	0,93	0,21	0,18	0,88
Under	Préalable	0,26	0,29	0,87	0,22	0,28	0,91	0,20	0,19	0,87
Both	Préalable	0,29	0,2	0,88	0,24	0,14	0,93	0,23	0,24	0,88

TABLE 3.13 – Indicateurs de performance des modèles avec sélection des variables avant rééchantillonnage

La méthode de correction des poids semble rendre le modèle plus performant que celui de l'ajustement préalable. Pour les piratages, la technique *Both sampling* est la plus performante. Il semble qu'aucune technique de rééchantillonnage n'améliore les résultats pour les piratages de prestataires et les divulgations/accès non autorisés.

Données	Correction	Piratage			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,33	0,15	0,88	0,3	0,19	0,92	0,27	0,14	0,87
Over	Poids	0,32	0,16	0,88	0,3	0,16	0,92	0,27	0,14	0,87
under	Poids	0,11	0,03	0,68	0,13	0,02	0,78	0,08	0,01	0,69
Both	Poids	0,33	0,17	0,86	0,26	0,14	0,92	0,26	0,16	0,87

TABLE 3.14 – Indicateurs de performance des modèles avec sélection des variables après rééchantillonnage

Lorsque la sélection des variables influentes est réalisée après les techniques de rééchantillonnage, cela ne permet pas d'améliorer les modèles. Il est en effet généralement recommandé de sélectionner les variables avant rééchantillonnage afin d'éviter de fausser les résultats. Cependant, les résultats ci-dessus doivent être interprétés avec précaution car ils ont été calculés à partir de la base d'apprentissage. La validation croisée permet d'avoir une estimation du F1-Score et de l'AUC s'ils avaient été calculés sur une base test. La valeur de $K = 10$ est choisie pour diviser la base afin d'avoir une division suffisamment large.

Il faut noter que selon les divisions obtenues de la base de données, les résultats peuvent être différents. Ainsi, pour pouvoir comparer les modèles entre eux, ils sont estimés à partir des mêmes divisions de la base. Le F1 Score est une mesure qui dépend du niveau de seuil à partir duquel une probabilité est associée à une réponse positive. Une validation croisée du F1-Score est effectuée pour chaque seuil possible, entre 0 et 1 avec un pas de 0,01, et la meilleure estimation est retenue avec son seuil associé.

Données	Correction	Piratage			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,26	0,16	0,82	0,25	0,13	0,87	0,20	0,15	0,83
Over	Poids	0,26	0,15	0,82	0,26	0,13	0,87	0,19	0,15	0,83
Under	Poids	0,19	0,33	0,79	0,16	0,00	0,85	0,20	0,48	0,81
Both	Poids	0,35	0,19	0,87	0,25	0,17	0,92	0,24	0,16	0,88
Over	Préalable	0,22	0,16	0,81	0,23	0,13	0,85	0,19	0,18	0,83
Under	Préalable	0,18	0,27	0,80	0,17	0,09	0,85	0,19	0,22	0,82
Both	Préalable	0,29	0,20	0,88	0,21	0,10	0,92	0,20	0,20	0,88

TABLE 3.15 – Validation croisée des indicateurs de performance des modèles

En se basant sur les résultats de validation croisée, la technique de rééchantillonnage la plus efficace d'après les valeurs du F1-Score et de l'AUC obtenues est *Both sampling* avec correction par poids, combinée à une sélection des variables avant le rééchantillonnage pour les piratages et les divulgations. Dans les cas des piratages de prestataires, le F1 score estimé augmente légèrement avec le suréchantillonnage, et l'AUC augmente avec le *Both sampling*. Dans ce contexte, comme il n'y a pas d'amélioration significative, une régression logistique sans rééchantillonnage est effectuée pour modéliser la probabilité de piratages de prestataires.

3.6 Résultats des modèles de probabilité de violations de données personnelles

3.6.1 Présentation des résultats des modèles retenus

Afin de faciliter la lecture des résultats, les modalités des variables sont regroupées avec des modalités proches afin d'augmenter la significativité des modalités, et de s'assurer que la variation de la probabilité d'incident est cohérente selon les différentes modalités. Si cette variation n'est pas cohérente malgré les regroupements, la variable est retirée. Les résultats des modèles sont les suivants :

	Coefficient	Ecart type	t valeur	Pr(>|t|)
Intercept	-8,4115	0,6917	-12,16	0,0000
Etat moyennement touché	1,2091	0,2527	4,79	0,0000
Etat fortement touché	1,8864	0,3046	6,19	0,0000
Seul hôpital membre d'un système	1,0028	0,2886	3,47	0,0005
Non membre d'un système	1,0378	0,2521	4,12	0,0000
Employes par Journee(0,012 ;0,034]	0,4616	0,3081	1,50	0,1341
Employes par Journee(0,034 ;12]	1,2012	0,3438	3,49	0,0005
Actif Total(6e+07 ;1,67e+10]	0,6963	0,2879	2,42	0,0156
Taux Marge Globale(0,18 ;1,69]	-1,3571	0,5714	-2,37	0,0176
Taux d'Occupation(0,4 ;0,6]	0,5876	0,4501	1,31	0,1917
Taux d'Occupation(0,6 ;1,01]	1,1612	0,4234	2,74	0,0061
Accueille des internes	-0,4389	0,2575	-1,70	0,0884
Salaire Moyen(7,8e+04 ;3e+05]	0,3751	0,2073	1,81	0,0705
Ratio CA Sorties(2e+04 ;3,55e+07]	1,1269	0,4864	2,32	0,0206
Impayes(3,3e+07 ;4,24e+08]	1,0190	0,2885	3,53	0,0004
Part de Marché(0,0075 ;0,027]	0,1954	0,2759	0,71	0,4788
Part de Marché(0,027 ;1,01]	0,8551	0,3307	2,59	0,0098

TABLE 3.16 – Résultats régression logistique probabilité de piratage

	Coefficient	Ecart type	t valeur	Pr(>|t|)
Intercept	-6,9749	0,6919	-10,08	0,0000
Etat moyennement touché	1,6227	0,3398	4,78	0,0000
Etat fortement touché	2,5249	0,3754	6,73	0,0000
Taux de Marge Globale(0,08 ; 1,69]	-0,9948	0,3768	-2,64	0,0083
Employés par Lit(3,5 ; 7]	0,8258	0,5927	1,39	0,1635
Employés par Lit(7 ;30]	1,8926	0,5903	3,21	0,0013
Ratio Couts Charges(0,24 ;0,5]	0,3018	0,3357	0,90	0,3687
Ratio Couts Charges(0,5 ;5,01]	-0,9100	0,5464	-1,67	0,0958
Ratio Impayes Couts(0,07 ;1,59]	-0,7106	0,3429	-2,07	0,0383
Impayes(6e+06 ;4,24e+08]	0,7938	0,3314	2,40	0,0166
Seul hôpital membre du système	0,6935	0,3425	2,02	0,0429
Non membre d'un système	0,4508	0,3449	1,31	0,1912
Employés par Journée(0,018 ;12]	-0,4241	0,3312	-1,28	0,2004
Liquidités(5e+06 ;3,3e+07]	0,2950	0,3442	0,86	0,3914
Liquidités(3,3e+07 ;4,08e+09]	1,1717	0,3615	3,24	0,0012

TABLE 3.17 – Résultats régression logistique probabilité de piratages de prestataires

	Coefficient	Ecart Type	t valeur	Pr(>|t|)
Intercept	-6,0131	0,7440	-8,08	0,0000
Etat peu touché	-1,4501	0,4329	-3,35	0,0008
Etat fortement touché	1,2921	0,4530	2,85	0,0044
Employés par Lit(3,5 ;6,5]	0,5881	0,5311	1,11	0,2682
Employés par Lit(6,5 ;30]	1,1479	0,5394	2,13	0,0333
Taux d'Occupation(0,7 ;1,01]	0,4038	0,2806	1,44	0,1502
Impayés(8e+06 ;2,1e+07]	0,5428	0,3711	1,46	0,1437
Impayés(2,1e+07 ;4,24e+08]	1,2869	0,3584	3,59	0,0003
Ratio CA Sorties(2e+04 ;3,55e+07]	1,1811	0,6191	1,91	0,0565
Type de SoinsCancer/Enfants/Autre	0,6816	0,4419	1,54	0,1230
Type de SoinsLong Terme	-0,8585	1,0473	-0,82	0,4124
Type de SoinsPsy/Readaptation/Reli	-2,0343	1,2907	-1,58	0,1150
Hopital Accès CritiqueOui	-1,0791	0,5421	-1,99	0,0466

TABLE 3.18 – Résultats régression logistique probabilité de divulgations

Les coefficients mis à l'exponentielle représentent les rapports des cotes associés à chaque modalité ou chaque variable pour la hausse d'une unité de chaque variable. Lorsque le rapport des cotes est supérieur à 1, ce qui correspond à un coefficient de la régression logistique supérieur à 0, cela signifie que la variable a un effet positif sur la probabilité d'incident. Inversement, s'il est inférieur à 1, l'effet est négatif.

La représentation graphique suivante permet de visualiser l'effet des différentes modalités.

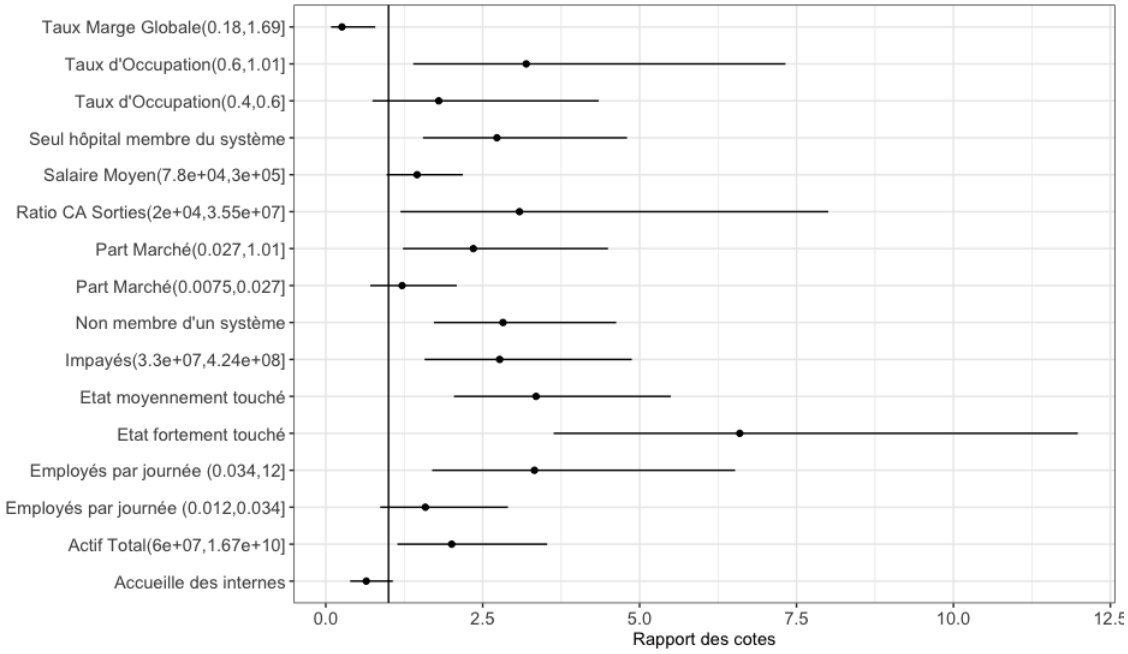


FIGURE 3.21 – Rapports des cotes du modèle des piratages

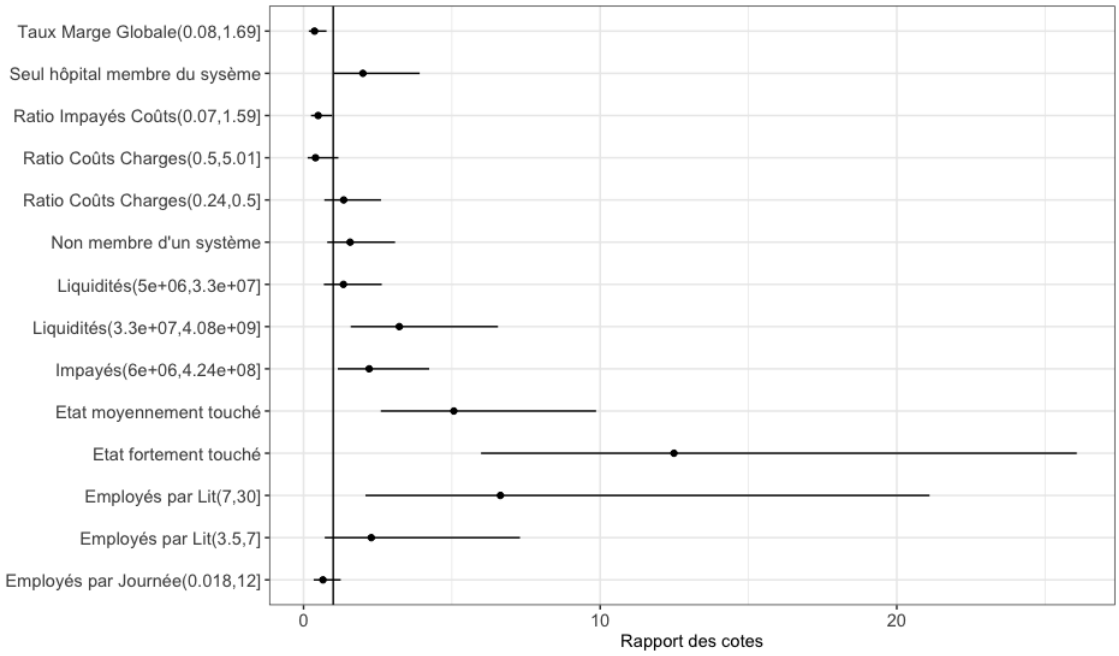


FIGURE 3.22 – Rapports des cotes piratages prestataires

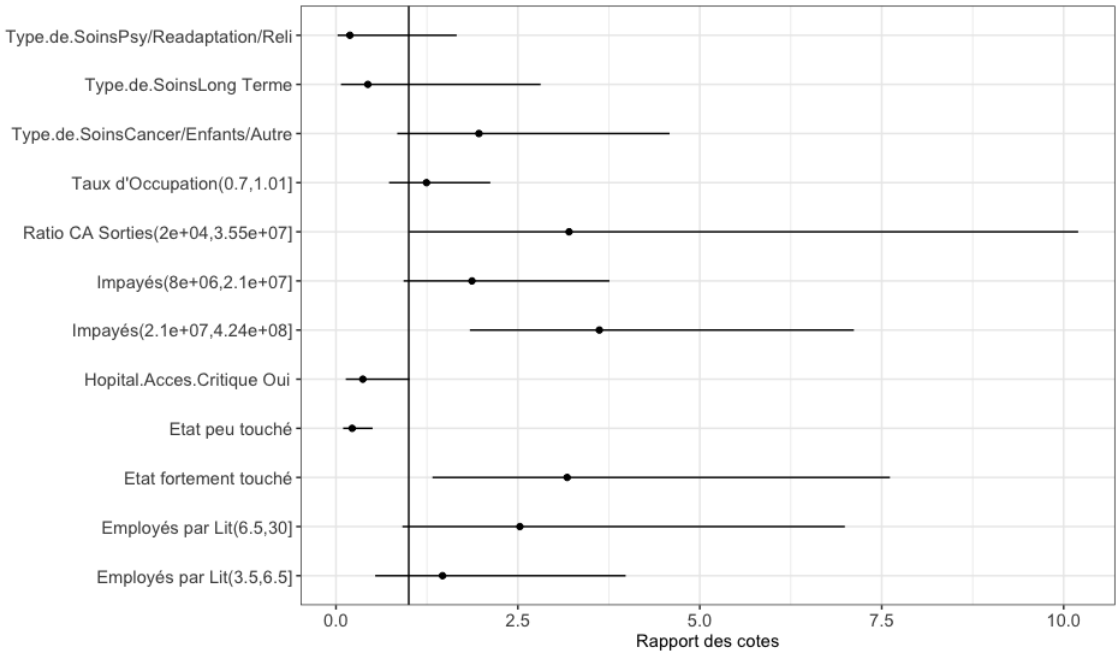


FIGURE 3.23 – Rapports des cotes du modèle des divulgations

Ainsi, quelques exemples d'interprétation sont donnés. Avoir une marge globale importante réduit la probabilité de piratage. Au contraire, avoir un taux d'occupation important, une forte part de marché, ne pas être membre d'un groupe ou être le seul hôpital d'un groupe augmente cette probabilité. Pour les piratages de prestataires et les divulgations, la probabilité est fortement augmentée s'il y a beaucoup d'employés par rapport au nombre de lits.

Le test d'Hosmer et Lemeshow est ensuite réalisé. Les résultats valident les modèles et sont résumés ci-dessous. La p valeur est dans chaque cas supérieure à 0,05, ce qui valide les modèles.

Catégorie d'incident	X2	df	p valeur
Piratages	4,3	8	0,8
Piratagess prestataires	13	8	0,1
Divulgations	13	8	0,1

TABLE 3.19 – Test d'Hosmer et Lemeshow

Le test de Wald est réalisé afin de déterminer s'il existe au moins un coefficient statistiquement différent de 0. Les p valeurs obtenues sont nulles et indiquent qu'il est possible de rejeter l'hypothèse nulle qui est que tous les coefficients valent 0.

Catégorie d'incident	X2	df	p valeur
Piratages	109,1	16	0
Piratages prestataires	160,3	14	0
Divulgations	233,6	47	0

TABLE 3.20 – Test de Wald

3.6.2 Annualisation de la probabilité selon l'évolution temporelle du nombre de déclarations des hôpitaux américains

L'objectif est d'estimer la probabilité annuelle pour un hôpital de déclarer au moins une violation de données.

Après segmentation selon les types d'incident décrits ci-dessus, il y a peu de données sur un an. Par exemple, 26 hôpitaux ont déclaré un incident de piratage en 2021. Les modèles sont donc calibrés sur une période élargie. Les règles de déclarations ayant été modifiées et finalisées en 2013, chaque modèle estime la probabilité pour un hôpital de connaître un incident entre 2014 et 2021.

Il s'agit ensuite de récupérer la probabilité de 2021 à partir de la probabilité estimée par le modèle sur la période 2014 à 2021.

Le graphique suivant illustre l'évolution de la probabilité annuelle moyenne segmentée par type d'incident :

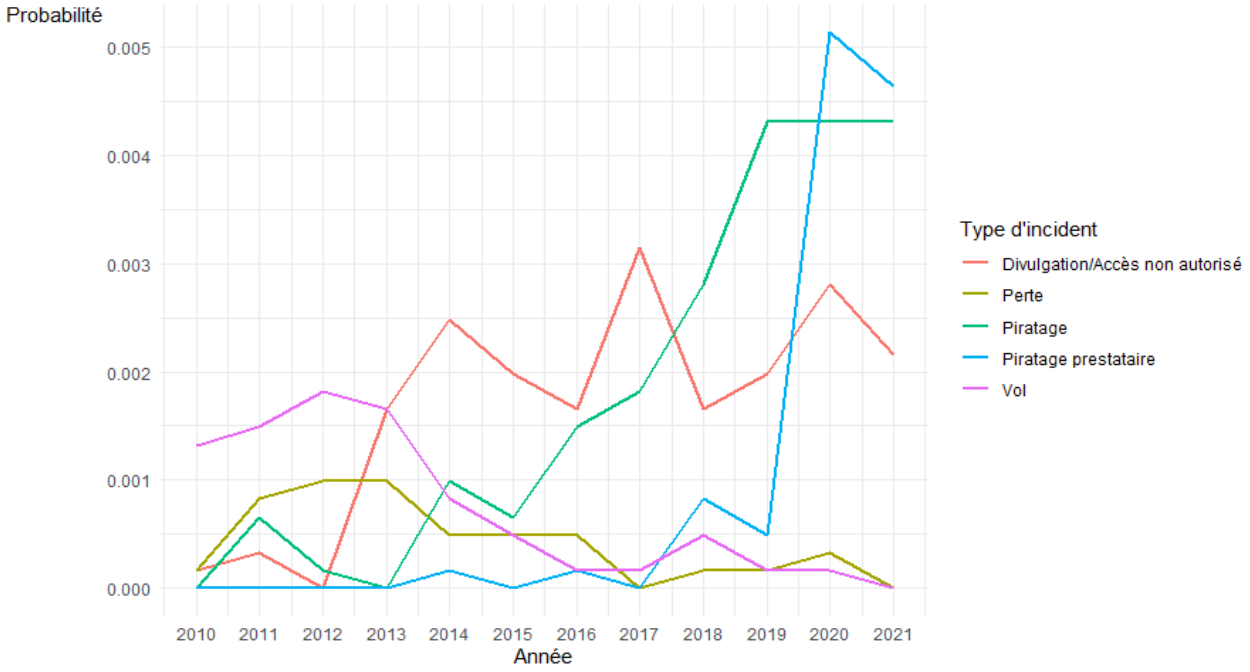


FIGURE 3.24 – Probabilité annuelle de déclaration des hôpitaux par catégorie d'incident

La tendance de forte hausse des piratages et baisse des vols observée sur le graphique 2.1 est retrouvée ici.

La probabilité de déclarer un incident entre 2014 et 2021 peut être décomposée par une somme de probabilités d'évènements disjoints correspondant aux probabilités de déclarer au moins un évènement chaque année. En effet, les établissements déclarant plusieurs incidents du même type à des années différentes sont rares. Cela peut s'expliquer par le fait qu'après un incident, l'*OCR* fournit des recommandations ou obligations aux entreprises concernant la sécurité informatique et la protection des données personnelles.

Soit I_t l'évènement qui vaut 1 si un incident a lieu en t, 0 sinon, et I l'évènement qui vaut 1 si un incident a lieu sur la période étudiée, 0 sinon.

$$P(\bigcup_{t=2014}^{2021} I_t|X) = \sum_{t=2014}^{2021} P(I_t|X)$$

Chaque probabilité annuelle de 2014 à 2020 est ensuite exprimée en fonction de celle de 2021, afin d'estimer la probabilité de 2021 en prenant en en compte l'évolution des probabilités dans le temps.

Une étude est réalisée ci-dessous pour étudier l'évolution annuelle de la probabilité de survenance de chaque catégorie d'incident. Cela permet de déduire la probabilité estimée annuelle en 2021, puis celle de 2022 peut être projetée de différentes manières selon l'hypothèse d'évolution future, par exemple une stagnation ou une hausse linéaire qui se poursuit.

Pour les piratages, une régression linéaire de la probabilité annuelle est ajustée en fonction de l'année. Les résultats et le tracé de la droite estimée sont présentés ci-dessous :

	Estimate	*Std. Error*	*t value*	$\Pr(>\|t\|)$
(*Intercept*)	-1.2172	0.1585	-7.68	0.0003
Année	0.0006	0.0001	7.69	0.0003

TABLE 3.21 – Régression linéaire de la probabilité de piratage annuelle

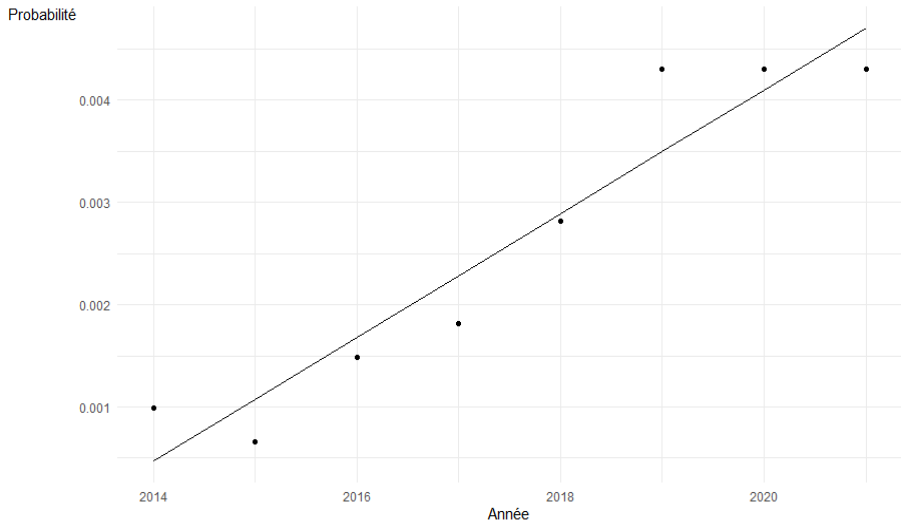


FIGURE 3.25 – Probabilité annuelle de piratage en fonction du temps

Le coefficient directeur de la droite estimé vaut $a = 0,0006$. Il est possible d'en déduire que la probabilité de piratage augmente de 0,0006 chaque année.

Ainsi,

$$P\left(I_{2021}\right) = a \times 2021 + \ \mathit{Intercept} = P\left(I_{2020}\right) + a = .. = P\left(I_{2014}\right) + 7a$$

Il est possible d'en déduire :

$$\begin{aligned} E\left(1_{I_{2021}}\right) &= a + E\left(1_{I_{2020}}\right) \\ E\left(E\left(1_{I_{2021}}|X\right)\right) &= E\left(a + E\left(1_{I_{2020}}|X\right)\right) \\ E\left(1_{I_{2021}}|X\right) &= a + E\left(1_{I_{2020}}|X\right) \end{aligned}$$

Cela permet d'écrire que

$$\begin{aligned} P\left(I|X\right) &= P\left(2021|X\right) + \sum_{k=1}^{7}\left(P\left(2021|X\right) - ka\right) \\ &= 8P(2021|X) - 28a \end{aligned}$$

Donc

$$P\left(I_{2021}|X\right) = \left(P\left(I|X\right) + 28a\right)/8 \tag{3.1}$$

De la même manière, les probabilités annuelles de vol peuvent être estimées par une régression linéaire. La tendance est décroissante, la probabilité de vol baisse d'environ 0,000085 par an. Voici les résultats du modèle et le tracé de la droite estimée :

	Estimate	*Std. Error*	*t value*	Pr(> |t|)
(*Intercept*)	0.1712	0.0590	2.90	0.0273
Année	-0.0001	0.0000	-2.90	0.0274

TABLE 3.22 – Régression linéaire de la probabilité de vol annuelle

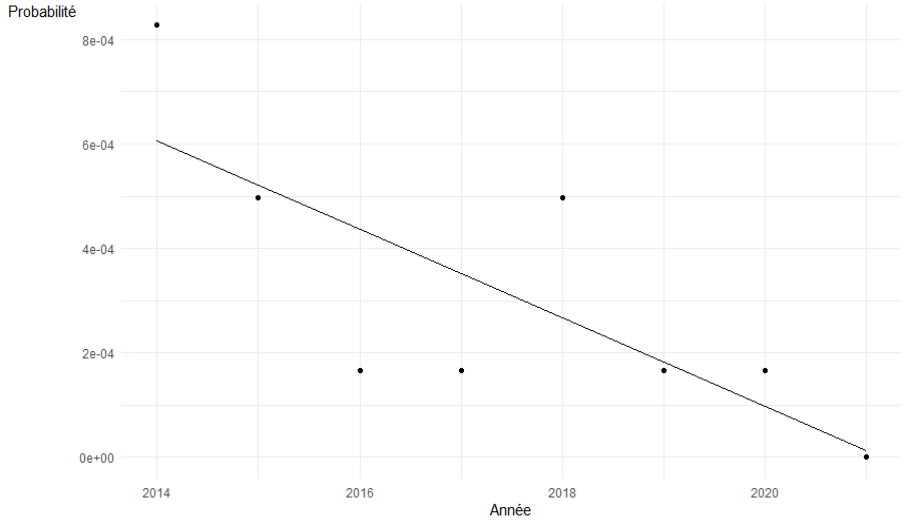


FIGURE 3.26 – Probabilité annuelle de vol en fonction du temps

Ainsi l'équation 3.1 peut être appliquée ici avec $a = -0,000085$. Par prudence, il est cependant choisi de considérer une probabilité annuelle constante, et donc la probabilité sera divisée par 8 afin d'obtenir une probabilité annuelle.

Pour les piratages des prestataires, une régression polynomiale avec un polynôme d'ordre 2 peut être ajustée. Les coefficients de la régression sont significatifs selon le test de Wald avec un seuil de 1% :

	Estimate	*Std. Error*	*t value*	Pr(> |t|)
(*Intercept*)	788.1546	332.4078	2.37	0.0639
Année	-0.7820	0.3295	-2.37	0.0637
Année^2	0.0002	0.0001	2.38	0.0635

TABLE 3.23 – Régression polynomiale de la probabilité de piratage des prestataires annuelle

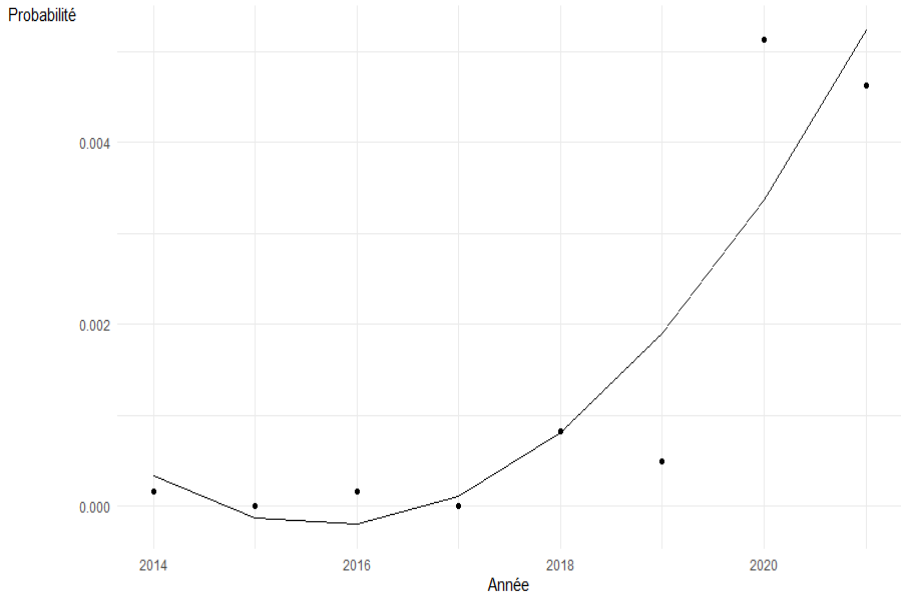


FIGURE 3.27 – Probabilité annuelle de piratage de prestataires en fonction du temps

L'année 2020 semble marquer une rupture, les piratages de prestataires étaient des évènements rares et constituent maintenant un risque important. Le choix est fait de considérer que les probabilités globales d'occurrence sont nulles entre 2014 et 2017, égales entre 2018 et 2019, puis que celles de 2021 et 2020 sont équivalentes. Graphiquement, il est possible d'estimer que la probabilité de 2021 est 5 fois supérieure à celle de 2019. Ainsi, cela revient à dire que

$$\begin{aligned} P_{globale} &= P_{2021} + P_{2020} + P_{2019} + P_{2018} \\ &= P_{2021} + P_{2021} + P_{2021}/5 + P_{2021}/5 \\ &= 12P_{2021}/5 \end{aligned}$$

ce qui revient à multiplier la probabilité de piratages de prestataires par $5/12$ pour obtenir une probabilité annuelle.

De fortes variations sont observées pour les divulgations / accès non autorisé et pertes d'outils informatiques dans le graphique 4.3. Les probabilités sont considérées constantes dans le temps pour ces catégories. Ainsi, afin de récupérer les probabilités annuelles, les probabilités modélisées sont divisées par 8.

Le tableau suivant résume les coefficients utilisés pour annualiser chaque probabilité estimée, avec p la probabilité modélisée sur toute la période :

Piratage	Piratage prestataire	Divulgation	Perte	Vol
$(p + 0,0168)/8$	$5p/12$	$p/8$	$p/8$	$p/8$

TABLE 3.24 – Annualisation de la probabilité

3.6.3 Ajustement de la fréquence liée au seuil de déclaration des données

Après annualisation de la probabilité, il est nécessaire de l'ajuster selon le seuil. En effet, les données utilisées pour la modélisation portent sur les incidents de plus de 500 individus. Il est choisi d'étudier le nombre d'incidents de moins de 500 individus afin d'ajuster la probabilité estimée pour prendre en compte tous les incidents. Cette probabilité est assimilable à une fréquence espérée, car il y a très peu d'établissements ayant déclaré plusieurs incidents dans les données ayant servi à la construction du modèle, ce qui n'est pas forcément le cas dans les déclarations de moins de 500 individus.

L'étude *Annual Report to Congress on Breaches of Unsecured Protected Health Information For Calendar Year 2020* publiée par *HHS* présente des statistiques globales et d'autres informations sur les violations de données de plus de 500 individus puis celles de moins de 500 individus.

La méthode d'ajustement de la probabilité de survenance est ici détaillée pour les piratages, sans distinguer les piratages de l'hôpital et les piratages du prestataire car la variable Présence d'un prestataire n'est pas utilisée dans l'étude. Elle révèle les statistiques suivantes concernant les incidents au-dessus du seuil :

— 77% des déclarations de 2020 sont celles d'établissements de soins
— 444 incidents de piratages ont été déclarés en 2020

Les informations ci-dessous sont énoncées concernant les incidents de moins de 500 individus.

— 90% des déclarations de 2020 sont celles d'établissements de soins
— 665 incidents de piratages ont été déclarés en 2020

Ainsi, les piratages dépassant le seuil de 500 individus représentent 40% de tous les piratages déclarés. Le pourcentage de piratages d'établissements de soins au-dessus du seuil par rapport à tous les piratages peut être estimé ainsi :

$$\frac{(444 \times 0,77)}{(665 \times 0,9 + 444 \times 0,77)} = 0,36$$

Donc, pour obtenir la fréquence totale des piratages d'établissements de soins à partir de la fréquence d'incidents au-dessus du seuil, il suffit de la multiplier par $1/0,36 = 2,8$

Le même coefficient sera utilisé pour ajuster la fréquence des hôpitaux, en faisant l'hypothèse que la proportion d'hôpitaux au sein des déclarations des établissements de santé au-dessus et en-dessous du seuil est la même. En 2020, 57 hôpitaux ont réalisé 57 déclarations d'incidents de piratage. La probabilité annuelle ou la fréquence de déclaration d'incidents de plus de 500 individus vaut donc :

$$\frac{57}{6045} = 0,95\%$$

La notation x désigne la proportion des hôpitaux au sein des déclarations des établissements de soins, identique pour les déclarations au-dessus et en-dessous du seuil. Ainsi, x vaut

$$x = 57/(444 \times 0,77) = 0,17$$

La fréquence de déclaration d'un incident de moins de 500 individus pour un hôpital vaut donc

$$\frac{0,9 \times 665 \times x}{6045} = 1,7\%$$

La fréquence peut donc être estimée en multipliant la probabilité annuelle au-dessus du seuil par :

$$\frac{0,95\% + 1,7\%}{0,95\%} = 2,8$$

De la même manière, les coefficients pour les autres catégories d'incidents sont les suivants.

	Piratage	Divulgation	Vol	Perte
Nombre > 500	444	148	36	16
Nombre < 500	665	61 973	1 038	2 662
Coefficient	2,8	490	35	196

TABLE 3.25 – Coefficients d'ajustement liés au seuil de déclaration

Pour les divulgations, seule une infime partie des incidents est connue. Multiplier par 490 la fréquence paraît cependant très excessif car la plupart des incidents ne mènent pas forcément à un sinistre. L'ajustement de la fréquence nécessitera des hypothèses qui seront détaillées dans le chapitre 4.

3.7 Identification des scénarios cités dans la variable Description

La variable Description permet de mieux comprendre ce que désignent les différents types d'incidents. Ainsi, les vols et pertes sont uniquement associés à des pertes ou vols d'objets comme un ordinateur ou une clé USB. Les divulgations ou accès non autorisés sont très majoritairement issues d'erreurs ou d'actes malveillants d'employés.

Ces informations ont été présentées aux experts en risque Cyber afin de leur apporter un maximum d'informations pour leur permettre d'estimer un coût financier lié aux incidents. Certaines informations ont été utilisées pour la construction de la matrice de coûts qui sera présentée par la suite.

En particulier, les types d'attaques les plus fréquemment cités menant à un piratage de l'hôpital ou de son prestataire sont les suivants :

— L'hameçonnage
— Le rançongiciel
— Le logiciel malveillant

Les formes d'incidents les plus fréquentes au sein de la catégorie Divulgation / Accès non autorisé sont les suivantes :

— L'erreur humaine (mail envoyé au mauvais destinataire, exposition des données par erreur sur Internet, ...)
— L'accès non autorisé à des données par un employé ou ancien employé
— La négligence (utilisation des données de manière non sécurisée)

Également, grâce à la variable Localisation de l'incident, il est possible de voir que les vols et pertes peuvent concerner des ordinateurs de bureau, ordinateurs portables et autres objets informatiques comme des disques durs et des clés USB.

Pour les piratages, les piratages de prestataires, et les divulgations/accès non autorisés subis par les hôpitaux, les proportions de chaque évènement listé ci-dessus sont calculées en comptant le nombre d'occurrences de l'évènement divisé par le nombre de lignes où la variable Description est présente et mentionne ce type d'informations.

	Piratages	Piratages prestataires	Divulgations
Hameçonnage	0,81	0,11	0
Rançongiciel	0,10	0,89	0
Logiciel malveillant	0,09	0	0
Erreur humaine	0	0	0,43
Accès non autorisé par un employé	0	0	0,4
Négligence	0	0	0,17

TABLE 3.26 – Proportions des évènements les plus fréquents au sein de chaque type d'incident

3.8 Modélisation du nombre d'individus affectés par un incident Cyber

Le nombre d'individus affectés est une information qui peut donner des indications sur l'ampleur d'un incident et donc son coût. Il sera utilisé pour déterminer certains coûts.

La distribution semble être différente pour chaque catégorie d'incident, avec une ressemblance pour les piratages et piratages de prestataires. De plus, le nombre moyen d'individus affectés varie selon l'incident : il est de 36 635 pour les piratages, 30 618 pour les piratages de prestataires, et 5 739 pour les divulgations.

Un ajustement d'une loi de probabilité est réalisé pour chaque incident au logarithme du nombre d'individus affectés par les incidents subis par les hôpitaux américains. Une loi normale de moyenne 8,9 et d'écart type 1,7 est ajustée pour les piratages, une loi normale de paramètre 9,1 et 1,6 pour les piratages de prestataires, ainsi qu'une loi de Gumbel de position 7,2 et d'échelle 0,86 pour les divulgations.

Les graphiques ci-dessous sont la représentation simultanée de la densité empirique et la densité théorique, ainsi que les quantiles de la distribution théorique et empirique représentés l'un par rapport à l'autre. Ils permettent d'étudier la qualité de l'ajustement :

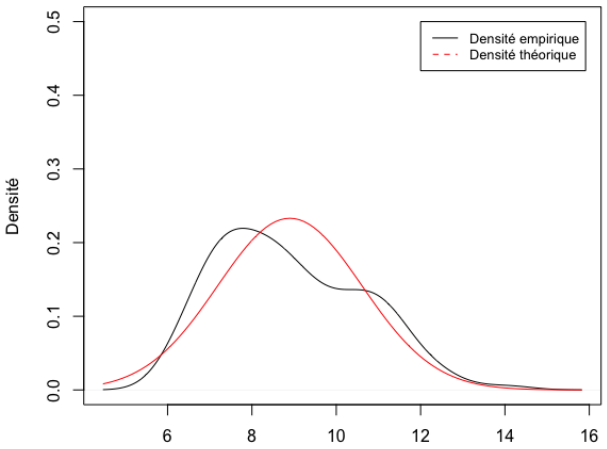


FIGURE 3.28 – Densité log(nombre d'individus affectés) par un piratage

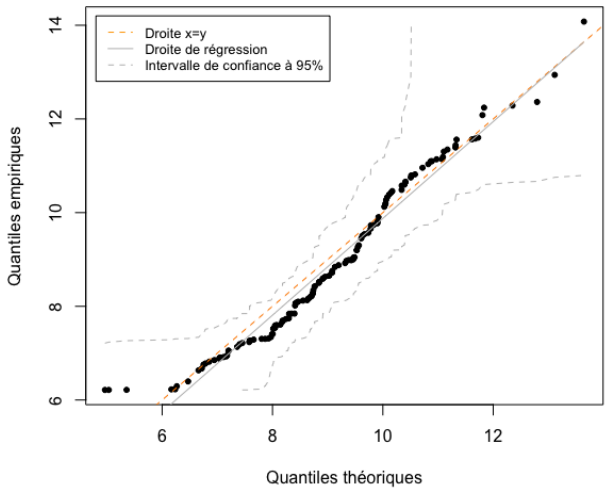


FIGURE 3.29 – QQ plot log(nombre d'individus affectés) par un piratage

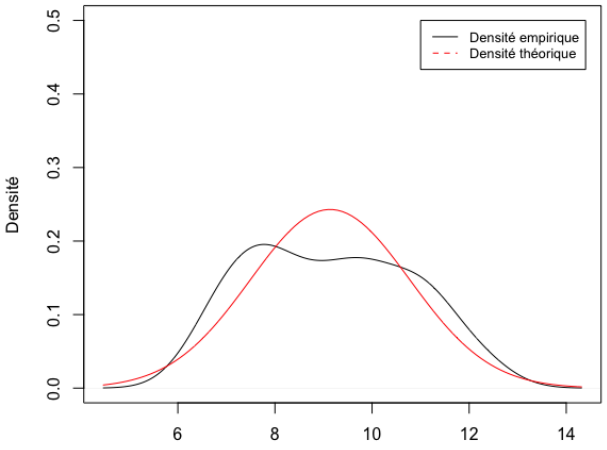


FIGURE 3.30 – Densité log(nombre d'individus affectés) par un piratage prestataire

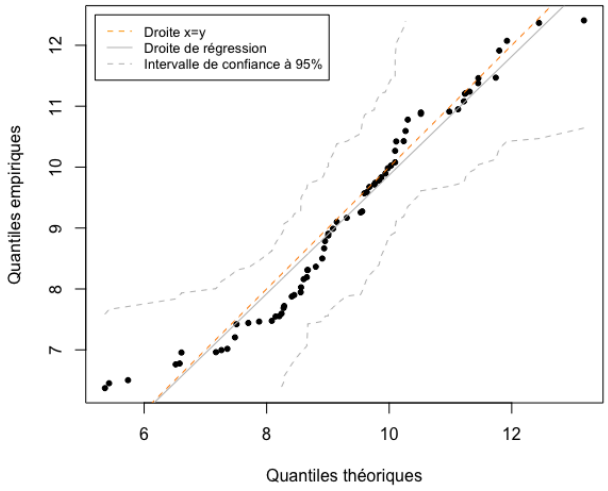


FIGURE 3.31 – QQ plot log(nombre d'individus affectés) par un piratage prestataire

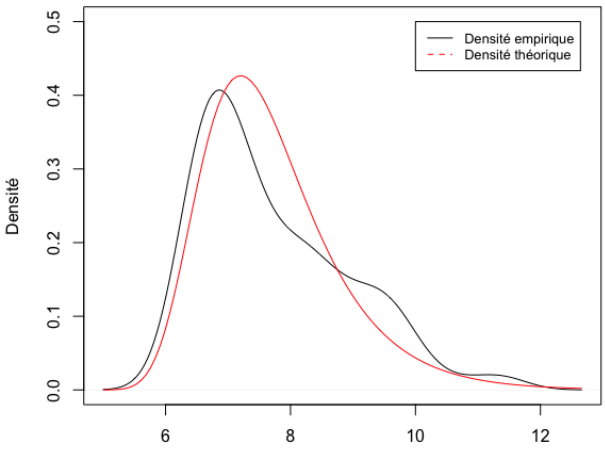


FIGURE 3.32 – Densité log(nombre d'individus affectés) par une divulgation/un accès non autorisé

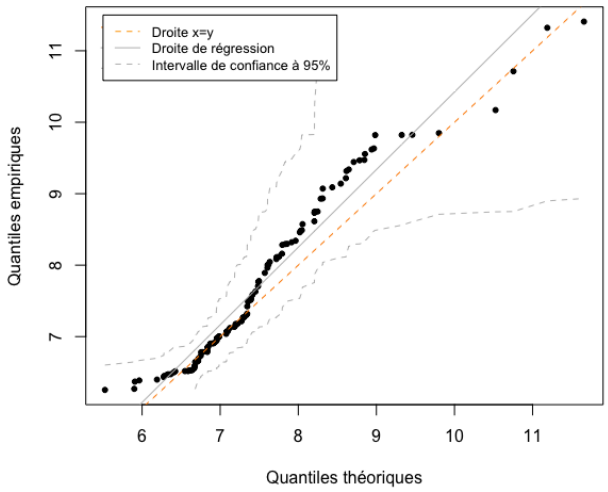


FIGURE 3.33 – QQ plot log(nombre d'individus affectés) par une divulgation/un accès non autorisé

Les courbes des densités empiriques et théoriques sont relativement proches, et les points dessinés sur les diagrammes quantile-quantile sont proches d'une droite.

Le test de Kolmogorov-Smirnov dont le fonctionnement est rappelé en annexe D permet de valider l'ajustement des lois. Il est appliqué sur l'échantillon des données, ainsi que sur un échantillon de données simulées depuis la loi estimée. Les résultats de ce test sont présentés ci-dessous :

	D	p valeur
Piratages	0,11	0,4
Piratages prestataires	0,14	0,5
Divulgations	0,12	0,4

TABLE 3.27 – Résultats tests de Kolmogorov-Smirnov

Le seuil de 500 individus a un impact sur la valeur des paramètres de la distribution estimée. En effet, les données observées correspondent à une distribution modifiée qui s'exprime ainsi :

$$\tilde{f}_{\theta|H}(x) = \frac{f_\theta(x)}{P(X > H)} \mathbb{1}_{x>H}$$

avec X la variable modélisée, $\tilde{f}$ la distribution modifiée par l'effet du seuil H et f la distribution si on observait toutes les données.

D'après la méthode des moments, on a

$$\mathbb{E}[X^p | X > H] = \int_{-\infty}^{\infty} x^p \tilde{f}_{\theta|H}(x) \mathrm{d}x$$

Une résolution numérique a été implémentée pour approximer les paramètres de la loi ajustée. Les

résultats des paramètres estimés pour chaque loi sont résumés dans le tableau suivant, avant ajustement par le seuil et après ajustement.

Incident	Loi modélisée	Moyenne	Ecart-type	Moyenne ajustée	Ecart-type ajusté
Piratages	Normale	8,9	1,7	8,2	2,2
Piratages prestataires	Normale	9,1	1,6	8,4	2,1

TABLE 3.28 – Ajustement des paramètres des lois normales estimées

Incident	Loi modélisée	Localisation	Echelle	Localisation ajustée	Echelle ajustée
Divulgations	Gumbel	7,2	0,86	7	0,9

TABLE 3.29 – Ajustement des paramètres de la loi de Gumbel estimée

En résumé, une approche de tarification inspirée du modèle fréquence × coût est adoptée. La probabilité, assimilable à une fréquence, est d'abord estimée à l'aide d'un modèle de régression logistique, avec application de rééchantillonnage pour faire face au déséquilibre des classes. Dans le chapitre 4, les coûts et les probabilités de déclenchement des garanties seront estimés par avis d'expert.

Dans cette partie, une modélisation américaine est effectuée, afin d'en apprendre plus sur le risque Cyber et d'identifier les profils d'établissements à risque d'incident. Une analyse bivariée est effectuée afin de ne conserver que les variables qui ne sont pas fortement liées entre elles. Les variables quantitatives sont ensuite discrétisées selon la mesure *WOE*, et les modalités de certaines variables qualitatives sont regroupées.

Le choix de la fonction de lien logit est justifié par l'interprétabilité des résultats. Les variables influentes sont ensuite sélectionnées par analyse de variance de type II, avant et après les différentes techniques de rééchantillonnage. La performance des modèles est comparée sur la base du F1-Score maximal possible selon les différents seuils d'affectation de réponse positive, ainsi que l'AUC. La technique de rééchantillonnage combinant sur-échantillonnage et sous-échantillonnage appliquée après sélection des variables semble ainsi être la plus efficace pour les piratages et divulgations, et une régression sans rééchantillonnage est effectuée pour modéliser la probabilité de piratage de prestataires.

Les modèles estimés permettent de mettre en évidence que la probabilité d'incident est plus élevée pour certaines catégories d'établissements, comme ceux dont le taux d'occupation est fort ou qui sont les seuls hôpitaux membres d'un système. Également, l'évolution temporelle de chaque incident est analysée puis prise en compte afin d'annualiser la probabilité modélisée sur toute la période. Une modélisation par ajustement de loi est effectuée sur le logarithme des individus affectés. La probabilité et le nombre d'individus estimés sont ajustés pour prendre en compte l'effet du seuil de déclaration des données.

Chapitre 4

Tarification dans le contexte français

Dans le chapitre précédent, les modèles ont permis de mettre en évidence certains profils plus à risque d'incident Cyber. Cependant, les variables utilisées ne sont pas toutes disponibles pour les sociétaires de Relyens. Afin de proposer une tarification adaptée au contexte du groupe, une seconde modélisation est réalisée en ne sélectionnant que des informations exploitables.

4.1 Sélection des variables disponibles pour les sociétaires de l'entreprise

De nombreuses variables présentées dans les modèles précédents ne sont pas disponibles pour un établissement français, ou bien pas comparables.

Afin de déterminer quelles variables sont utilisables en France, plusieurs sources d'information sont prises en compte. D'abord, Relyens détient des informations sur ses sociétaires demandées au titre de l'assurance Cyber ou bien d'autres assurances. De plus, la DRESS (direction de la Recherche, des Études, de l'Évaluation et des Statistiques) publie chaque année une base statistique sur les établissements de santé qui permet d'avoir des informations comme le nombre de journées de patient. Enfin certaines informations sont facilement trouvées sur internet comme le type de soins de l'établissement.

Également, les normes comptables américaines et de manière plus générale le contexte économique et culturel sont différents entre les Etats-Unis et la France, les variables conservées sont donc parfois proches mais non totalement exactes aux informations en France.

Ainsi, il sera considéré que les variables disponibles en France sont les suivantes :

— Nombre de lits
— Nombre d'employés ETP
— Chiffre d'Affaires
— Résultat net assimilé à la marge brute
— Membre d'un système associé à Membre d'un groupe de type GHT ou GCS (Groupement de Coopération Sanitaire) qui sera comparé à membre d'un système
— Hôpital d'accès critique qui sera équivalent au statut d'hôpital de proximité.
— Certification données, comparable au dossier médical partagé [43]
— Logiciel externe données

43. Depuis 2021, il n'est plus possible de créer un dossier médical partagé

— Type de soins
— Type de contrôle
— Accueille des internes
— Service d'urgences
— Nombre de journées
— Nombre de journées-lits
— Nombre de sorties qui sera comparé au nombre de séjours.

Ainsi, il est possible d'intégrer dans le modèle également les variables ratio suivantes :

— Taux d'occupation
— Sorties par lit
— Durée moyenne de séjour
— Revenu par lit
— Taux de marge globale
— Employés par lit
— Employés par journée
— Ratio CA Sorties

La variable membre d'un système est conservée après comparaison des organisations su système de soins dans les deux pays. Les établissements de soins aux Etats-Unis sont souvent regroupés et gérés par une entreprise, les reliant ainsi à d'autres établissements de soins. 68% des hôpitaux sont regroupés en systèmes en 2020 [44]. D'après *AHRQ (Agency for Healthcare Research and Quality)*, un système est formé d'« au moins un hôpital et au moins un groupe de médecins fournissant des soins complets, et qui sont liés les uns aux autres et avec l'hôpital par le biais d'une propriété commune ou d'une gestion commune » [45]. Dans une certaine mesure, cette organisation peut être comparée avec les regroupements d'établissements de soins qui ont lieu en France comme les GHT, qui peuvent se traduire par des regroupements de dossiers ou bien de serveurs.

De plus, la variable Type de contrôle n'est pas retenue. La segmentation public, privé et non lucratif est présente à la fois en France et aux Etats-Unis, mais elle n'est probablement pas directement comparable.

De la même manière que précédemment, seules les variables dont la corrélation linéaire est inférieure à 0,6 sont retenues, et seules les variables qualitatives dont le coefficient de Cramer n'indique pas une valeur supérieure à 0,7 sont retenues.

Le tableau ci-dessous récapitule les variables conservées pour la modélisation d'un établissement en France, ainsi que les valeurs prises pour vérifier qu'elles sont comparables à celles d'un établissement français, après retraitement des valeurs aberrantes. Le passage du dollar à l'euro est considéré négligeable, compte tenu du contexte actuel de parité entre les deux devises.

44. *Fast Facts on U.S. Hospitals Infographics, AHA(American Hospital Association), 2022*
45. Définition traduite https ://www.ahrq.gov/chsp/data-resources/compendium.html

	Minimum	1er quartile	Moyenne	3e quartile	Maximum
Nombre de Lits	2	32	153	199	2 826
Sorties par Lit	0	12,6	30,6	46	100
Nombre d'employés ETP	5	131,2	852,9	945,8	26491,2
Employés par Lit	0	2,7	5,3	6,6	30
Revenu par Lit	0	431 869	1 163 303	1 487 074	40 000 000
Ratio CA Sorties	0	21 785	79 938	59 253	10 000 000
Employés par Journée	0	0,01	0,03	0,03	4,00
Resultat Net (en millions)	-1 098	-1	7	10	1 623
Taux d'Occupation	0	0,51	0,69	0,92	1,00
Chiffre d'Affaires (en millions)	0	22	195	206	6 216
Durée Moyenne Séjour	0,3	4,7	35,7	27,3	400,00
Taux Marge Globale	-2	-0,04	-0,01	0,10	1,68
Nombre d'Internes	0	0,00	22	0,00	1611
Nombre de Jours-Lits	730	11 680	55 796	72 855	1 031 490
Nombre Sorties	1	489	5 929	7 737	130 889
Nombre de Journées	11	10 669	44 999	55 501	816 698

TABLE 4.1 – Variables du modèle France

Ensuite, les variables sont également discrétisées de la même manière que précédemment afin d'améliorer la qualité des modèles.

4.2 Modélisation de la probabilité d'incident à partir des variables vision France

Pour chaque catégorie d'incident, la probabilité de survenue est modélisée avec les mêmes techniques de rééchantillonnage et de correction que celles utilisées dans le chapitre précédent.

Les variables influentes après sélection par le test de type II à partir des régressions logistiques sans application de méthode de rééchantillonnage sont les suivantes :

	LR Chisq	Df	Pr(>Chisq)
Taux Marge Globale	28.86	5	0.0000
Chiffre d'Affaires	22.16	4	0.0002
Ratio CA Sorties	20.00	3	0.0002
Employés par Journée	21.12	5	0.0008
Membre Système	9.86	1	0.0017
Taux d'Occupation	20.41	6	0.0023
Enseignement	8.69	1	0.0032
Hôpital d'Accès Critique	5.27	1	0.0216
Sorties par Lit	10.89	4	0.0278
Type de Soins	6.26	3	0.0997

TABLE 4.2 – Variables vision France influentes sur la probabilité de piratage

	LR Chisq	Df	Pr(>Chisq)
Taux Marge Globale	25.27	3	0.0000
Chiffre d'Affaires	28.35	3	0.0000
Employés par Lit	21.19	3	0.0001
Type de Contrôle	7.46	2	0.0240

TABLE 4.3 – Variables vision France influentes sur la probabilité de piratage de prestataires

	LR Chisq	Df	Pr(>Chisq)
Chiffre d'Affaires	55,78	4	0,0000
Nombre d'employés par lit	16,75	3	0,0008
Taux d'occupation	16,88	6	0,0097
Chiffre d'Affaires/Nombre de sorties	12,03	3	0,0073
Type de soins	7,35	3	0,0617
Taux de marge globale	12,17	7	0,0951

TABLE 4.4 – Variables vision France influentes sur la probabilité de divulgation

Les mêmes indicateurs sont calculés pour chaque modèle afin de déterminer le plus performant.

		Piratage			Piratages prestataires			Divulgations		
Données	Correction	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,21	0,14	0,82	0,17	0,09	0,86	0,22	0,12	0,84
Over	Poids	0,21	0,13	0,82	0,17	0,09	0,86	0,21	0,11	0,84
Under	Poids	0,19	0,10	0,81	0,17	0,16	0,85	0,15	0,11	0,83
Both	Poids	0,21	0,13	0,82	0,17	0,09	0,86	0,23	0,16	0,84
Over	Préalable	0,18	0,21	0,83	0,17	0,09	0,86	0,18	0,10	0,84
Under	Préalable	0,18	0,08	0,82	0,17	0,13	0,86	0,15	0,10	0,83
Both	Préalable	0,17	0,15	0,83	0,17	0,09	0,86	0,20	0,14	0,84

TABLE 4.5 – Indicateurs de performance des modèles vision France avec sélection des variables avant rééchantillonnage

Données	Correction	Piratages			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,21	0,14	0,82	0,17	0,09	0,86	0,22	0,12	0,84
Over	Poids	0,21	0,10	0,83	0,25	0,17	0,88	0,23	0,11	0,84
Under	Poids	0,05	0,01	0,55	0,05	0,17	0,65	0,08	0,01	0,69
Both	Poids	0,21	0,14	0,82	0,17	0,09	0,86	0,23	0,16	0,84

TABLE 4.6 – Indicateurs de performance des modèles vision France avec sélection des variables après rééchantillonnage

Données	Correction	Piratages			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,16	0,12	0,78	0,16	0,05	0,82	0,19	0,13	0,80
Over	Poids	0,18	0,15	0,78	0,15	0,07	0,82	0,18	0,12	0,81
Under	Poids	0,11	0,13	0,75	0,16	0,12	0,82	0,14	0,24	0,80
Both	Poids	0,20	0,13	0,82	0,17	0,08	0,85	0,22	0,16	0,85
Over	Préalable	0,16	0,15	0,77	0,16	0,01	0,82	0,17	0,10	0,81
Under	Préalable	0,12	0,09	0,77	0,15	0,08	0,81	0,15		0,81
Both	Préalable	0,16	0,18	0,83	0,16	0,08	0,85	0,21	0,14	0,85

TABLE 4.7 – Validation croisée des indicateurs de performance des modèles vision France, sélection des variables avant rééchantillonnage

Données	Correction	Piratages			Piratages prestataires			Divulgations		
		F1-Score	Seuil	AUC	F1-Score	Seuil	AUC	F1-Score	Seuil	AUC
Initiales		0,16	0,12	0,78	0,16	0,05	0,82	0,19	0,13	0,80
Over	Poids	0,16	0,13	0,77	0,19	0,12	0,81	0,17	0,08	0,80
Both	Poids	0,19	0,10	0,82	0,17	0,08	0,85	0,22	0,16	0,85

TABLE 4.8 – Validation croisée des indicateurs de performance des modèles vision France, sélection des variables après rééchantillonnage

Ainsi, la technique de rééchantillonnage *both sampling* avec sélection des variables avant rééchantillonnage est la plus performante pour modéliser la probabilité de divulgation. Concernant les piratages, les indicateurs calculés sur toute la base ne mettent pas en avant de technique de rééchantillonnage permettant d'améliorer la régression logistique, mais la validation croisée montre une hausse non négligeable du F1 Score et de l'AUC avec la méthode *both sampling*, cette méthode est donc retenue. Concernant la probabilité de piratage de prestataires, c'est le suréchantillonnage puis la sélection des variables influentes qui permet d'avoir le F1-score le plus élevé.

4.3 Modèles retenus dans le contexte du secteur de la santé en France

Comme lors du chapitre 3, certaines modalités de variables des modèles finaux sont regroupées afin d'augmenter le nombre de coefficients significatifs et de s'assurer que l'évolution de la probabilité selon les classes est cohérente.

Les résultats des modèles les plus performants selon la section précédente sont présentés ci-dessous.

	Coefficient	Ecart type	t value	Pr(>|t|)
(Intercept)	-4,9966	0,3886	-12,86	0,0000
Taux Marge Globale(0,04 ;0,18]	-0,4826	0,2153	-2,24	0,0250
Taux Marge Globale(0,18 ;1,69]	-1,8018	0,5882	-3,06	0,0022
Chiffre d'Affaires(6e+07 ;2,4e+08]	0,5029	0,3261	1,54	0,1230
Chiffre d'Affaires(2,4e+08 ;7,2e+08]	1,0925	0,4130	2,65	0,0082
Chiffre d'Affaires(7,2e+08 ;6,22e+09]	2,0914	0,5070	4,12	0,0000
Ratio CA Sorties(1,5e+04 ;2e+04]	-0,9043	0,5096	-1,77	0,0760
Ratio CA Sorties(6,5e+04 ;1e+07]	0,5727	0,2807	2,04	0,0414
Ratio CA Sorties[0,1 ;5e+04]	-1,6815	0,9674	-1,74	0,0822
Employés par Journée(0,024 ;0,034]	0,0821	0,3027	0,27	0,7862
Employés par Journée(0,034 ;4,01]	0,7732	0,2734	2,83	0,0047
Non membre d'un système	0,6256	0,2473	2,53	0,0114
Taux d'Occupation(0,6 ;1,01]	0,7619	0,2803	2,72	0,0066
Accueille des internes	-0,8216	0,2886	-2,85	0,0044
Hopital Accès Critique	-0,7294	0,3665	-1,99	0,0467
Sorties par Lit(30 ;100]	0,3393	0,3160	1,07	0,2830
Type de Soins Cancer/Enfants/Autre	0,7443	0,3812	1,95	0,0509
Type de Soins Long Terme	-2,0933	1,1320	-1,85	0,0645
Type de Soins Psychiatrique/Réadaptation/Religieux	-0,4708	0,4311	-1,09	0,2748

TABLE 4.9 – Résultats modèle vision France probabilité de piratage

	Coefficient	Ecart type	t valeur	Pr(>|t|)
(Intercept)	-5,8142	0,2905	-20,01	0,0000
Taux Marge Globale(0,08 ;0,18]	-1,3235	0,2731	-4,85	0,0000
Taux Marge Globale(0,18 ;1,69]	-2,2989	0,6579	-3,49	0,0005
Chiffre d'Affaires(2,8e+08 ;6,8e+08]	1,0432	0,2475	4,21	0,0000
Chiffre d'Affaires(6,8e+08 ;6,22e+09]	1,2678	0,2976	4,26	0,0000
Chiffre d'Affaires[0 ;2e+07]	-1,0494	0,4943	-2,12	0,0338
Type de SoinsCancer/Enfants/Autre	1,1737	0,2692	4,36	0,0000
Type de SoinsLong Terme	-13,7703	405,3999	-0,03	0,9729
Type de SoinsPsy/Readaptation/Reli	-1,0924	0,6550	-1,67	0,0954
Sorties par Lit(30 ;36]	-0,2992	0,3720	-0,80	0,4212
Sorties par Lit(36 ;48]	-0,2395	0,2803	-0,85	0,3929
Sorties par Lit(48 ;100]	-0,0070	0,2845	-0,02	0,9804
Taux d'Occupation(0,68 ;0,8]	0,2332	0,2804	0,83	0,4056
Taux d'Occupation(0,8 ;1,01]	1,0230	0,2546	4,02	0,0001
Resultat Net(4e+06 ;1,8e+07]	0,6763	0,2278	2,97	0,0030
Resultat Net(1,8e+07 ;1,62e+09]	0,4629	0,2631	1,76	0,0786
Employés par Lit(6 ;30]	0,7949	0,2354	3,38	0,0007
Employés par Journée(0,018 ;0,028]	0,1083	0,2692	0,40	0,6873
Employés par Journée(0,028 ;4,01]	0,5703	0,2875	1,98	0,0473

TABLE 4.10 – Résultats modèle vision France probabilité de piratage de prestataires

	Coefficient	Ecart type	t valeur	Pr(>|t|)
(Intercept)	-5,9780	0,5976	-10,00	0,0000
Chiffre d'Affaires(8e+07 ;2,6e+08]	0,8453	0,3614	2,34	0,0194
Chiffre d'Affaires(2,6e+08 ;6,6e+08]	1,6235	0,3494	4,65	0,0000
Chiffre d'Affaires(6,6e+08 ;6,22e+09]	2,2465	0,3539	6,35	0,0000
Employés par Lit(3,5 ;6,5]	0,3868	0,4444	0,87	0,3841
Employés par Lit(6,5 ;10]	0,4553	0,4739	0,96	0,3368
Employés par Lit(10 ;30]	1,1894	0,4794	2,48	0,0131
Taux d'Occupation(0,7 ;1,01]	0,1652	0,2353	0,70	0,4826
Ratio CA Sorties(2e+04 ;1e+07]	0,6346	0,5008	1,27	0,2051
Type de Soins Cancer/Enfants/Autre	0,4068	0,3774	1,08	0,2811
Type de Soins Long Terme	-0,1093	0,8502	-0,13	0,8977
Type de Soins Psychiatriques/Réadaptation/Religieux	-1,3422	1,0171	-1,32	0,1870
Taux Marge Globale(0,08 ;0,22]	-0,5267	0,2667	-1,98	0,0483
Taux Marge Globale(0,22 ;1,69]	-0,5669	0,5452	-1,04	0,2984

TABLE 4.11 – Résultats modèle vision France probabilité de divulgation

Les rapports des cotes sont comme précédemment illustrés graphiquement :

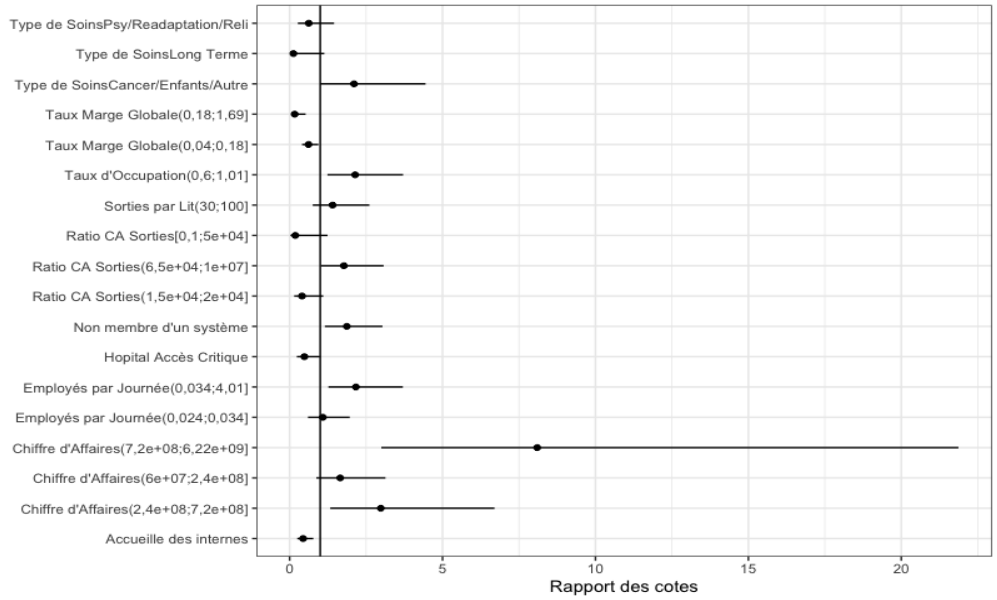


FIGURE 4.1 – Rapport des cotes piratages vision France

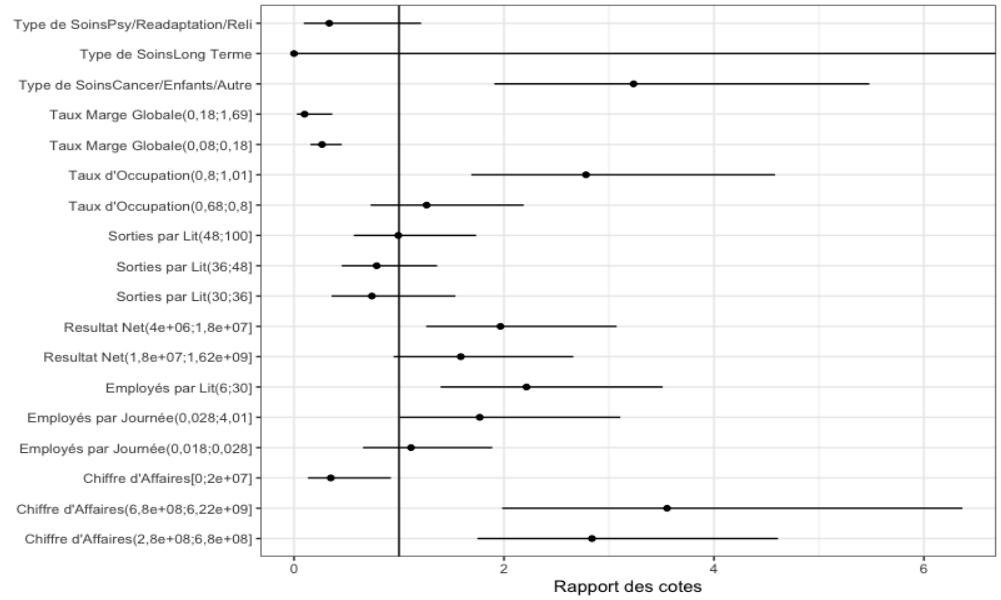


FIGURE 4.2 – Rapport des cotes piratages de prestataires vision France

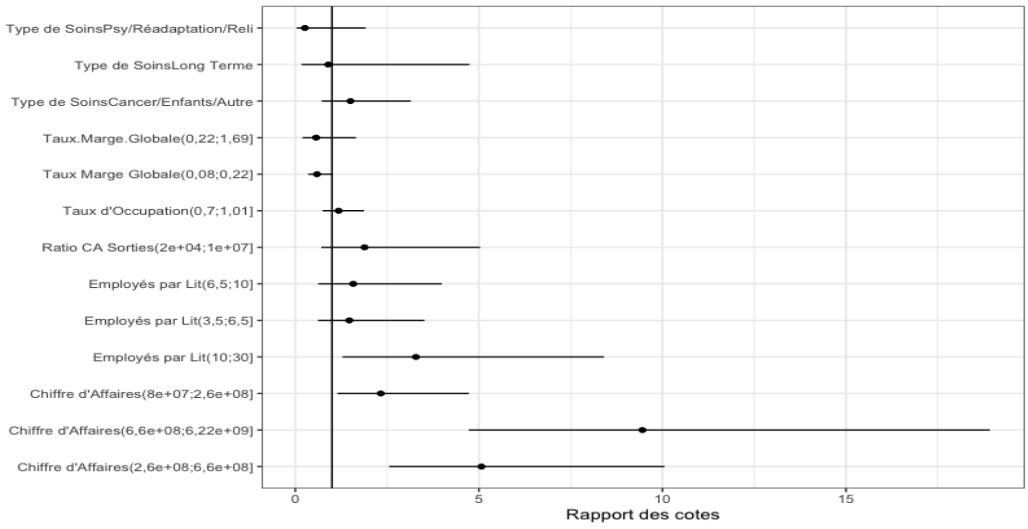


FIGURE 4.3 – Rapport des cotes divulgations vision France

Pour chaque type d'incident, le chiffre d'affaires est la variable la plus influente : un chiffre d'affaires élevé augmente la probabilité estimée. Le taux de marge globale réduit la probabilité. Un fort taux d'occupation est un facteur de risque, en particulier pour les piratages et piratages de prestataires. Il peut également être noté que la hausse de la variable Ratio CA Sorties augmente les probabilités de chaque incident, et celles de la variable Employés par journée augmente le risque de piratage et piratage de prestataires. Les effets des variables sont donc similaires entre les modèles, avec des impacts plus ou moins forts et quelques spécificités. Par exemple, le statut d'hôpital de proximité réduit la probabilité de piratage. Cela peut s'expliquer par le fait qu'il s'agisse d'hôpitaux plus petits et moins exposés. Enfin, le fait de ne pas être membre d'un groupe augmente la probabilité de piratage : il est possible de penser qu'un établissement seul ne dispose pas du même budget de sécurité informatique.

Les tests d'Hosmer et Lemeshow et de Wald valident les modèles.

Catégorie	X2	df	p valeur
Piratage	13	8	0,1
Piratage prestataires	3,9	8	0,9
Divulgations	3,7	8	0,9

TABLE 4.12 – Test Hosmer et Lemeshow

Catégorie	X2	df	p valeur
Piratage	122,6	18	0
Piratage prestataires	163,3	15	0
Divulgations	91,2	9	0

TABLE 4.13 – Test Wald

4.4 Ajustements de la fréquence modélisée

Chaque modèle estime la probabilité d'une catégorie d'incident en fonction des caractéristiques de l'établissement. Comme présenté dans la partie 3, la probabilité est annualisée puis ajustée selon le seuil, assimilée à une fréquence. Ensuite, cette fréquence peut être ajustée pour prendre en compte le contexte du monde de la santé en France.

4.4.1 Adaptation au marché français

Le rapport public 2021 de l'Observatoire des signalements d'incidents de sécurité des systèmes d'information pour le secteur santé donne des chiffres sur les incidents Cyber en France.

En 2021, il y a eu en France 733 déclarations par 582 structures, dont 479 déclarations réalisées par des établissements de santé. Le nombre de déclarations a doublé par rapport à celui de 2020. Il y a en France 3 036 établissements de santé. Ainsi, la fréquence annuelle de déclaration des incidents Cyber peut être estimée par

$$\frac{479}{3036} = 16\%.$$

Comme vu dans la section 3.6.3, il est considéré que les hôpitaux représentent 17% des déclarations, et $0,9$ correspond à la proportion de déclarations sous le seuil effectuées par des établissements de soins. L'étude *Annual Report to Congress on Breaches of Unsecured Protected Health Information For Calendar Year 2020* publiée par HHS indique que 67% des incidents sous le seuil concernent des données papier, il sera donc considéré qu'il y a 33% d'incidents Cyber sous le seuil pour les catégories, hormis le piratage car il est supposé que sa localisation ne peut pas être le papier. Également, il y a 61 973 divulgations, 1 038 vols, 2 662 pertes et 665 piratages de moins de 500 individus déclarés en 2020.

Il est donc possible d'estimer le nombre de violations de données personnelles de santé sous le seuil par :

$$665 \times 0,9 + 61973 \times 0,9 \times 0,33 + 1038 \times 0,9 \times 0,33 + 2662 \times 0,9 \times 0,33 = 20103$$

En considérant une proportion de 17% d'hôpitaux touchés, il y a donc 3418 déclarations d'hôpitaux

sous le seuil.

La fréquence moyenne d'incidents pour les hôpitaux, tout incident confondu tant qu'il est Cyber, est donc de

$$\frac{80 + 3418}{6045} = 0,58$$

Ainsi, la fréquence modélisée sera multipliée par 0,16/0,58.

4.5 Matrice de coûts des incidents Cyber

A partir des catégories d'incidents et des évènements cités dans la variable Description, dont les modalités obtenues sont listées en 2.2.2 un tableau a été créé pour classifier les incidents Cyber présents dans la base sinistres étudiée en scénarios.

Le produit d'assurance proposé par Relyens est constitué de différentes garanties. Certaines consistent en l'assurance de la responsabilité civile de l'établissement en cas d'atteinte au tiers à la suite d'un incident informatique. D'autres sont liées à la réparation des dommages causés par un incident à l'établissement.

A chaque type d'incident est associé un coût par garantie du contrat. Ce coût est déterminé par avis d'expert, selon les connaissances des experts en risque Cyber de l'entreprise, en estimant les conséquences les plus probables avec prudence. Il peut être forfaitaire, dépendre de caractéristiques de l'établissement, ou bien du nombre d'individus affectés par l'incident.

Pour chaque incident, une probabilité de déclenchement de garantie est déterminée. En effet, les incidents ne mènent pas forcément tous à des sinistres. Dans un premier temps, les incidents qui ne rentrent pas dans le cadre du contrat d'assurance ont été exclus de la modélisation de la fréquence d'incidents Cyber. Les modèles sont donc représentatifs de situations à risque, qui peuvent déclencher une ou plusieurs garanties du contrat d'assurance. Une des limites de ces données est que les conséquences des incidents ne sont pas connues. Supposer que les garanties sont déclenchées dans tous les cas est cependant très excessif. Par exemple, lorsque des données personnelles de santé sont divulguées, cela se traduit par un sinistre en France uniquement si un préjudice est subi par l'individu ou les individus affectés, et qu'une procédure judiciaire a lieu. Les probabilités de déclenchement sont estimées en observant les sinistres connus par le groupe, et par avis d'expert.

Dans un cas, le déclenchement d'une garantie était conditionné à la survenue d'un certain type d'évènement au sein de la catégorie d'incident. La proportion de cet évènement a donc été multipliée avec la probabilité de déclenchement de la garantie.

La matrice de coût est présentée ci-dessous. Pour des raisons de confidentialité, les garanties ne sont pas définies précisément, les formules permettant de calculer les coûts sont simplifiées, et les valeurs présentées ont été multipliées par un coefficient.

	Garantie dommages/assistance A	Garantie dommages/assistance B	Garantie dommages/assistance C	Garantie Responsabilité Civile
Déclenchement piratages	0,4	0,4	0,45	0,005
Déclenchement divulgations	0,1	0,1	0,1	0,005
Déclenchement pertes	0	0,05	0	0,005
Déclenchement vols	0	0,4	0	0,005
Coût Piratages/Divulgations	291 600	10 206 × de serveurs + 49 × Nombre d'ordinateurs	170 × Nombre de consultations urgentes + 80 × Nombre de consultations non urgentes + 0,5% du Chiffre d'affaires	49 × Nombre d'individus affectés + 1458 × Nombre d'employés
Coût Pertes/Vols		10 206 × Nombre de serveurs + 49 × Nombre d'ordinateurs		49 × Nombre d'individus affectés + 1458 × Nombre d'employés

TABLE 4.14 – Matrice de coûts par garantie

4.6 Construction du tarificateur

Le tarificateur est implémenté sous R et est à destination des actuaires de l'entreprise. Une zone permet d'abord d'indiquer les différentes informations de l'établissement, comme le chiffre d'affaires ou le nombre de lit. Les variables de ratio se calculent ensuite automatiquement, et les prédictions des probabilités sont réalisées à partir des modèles visions France pour chaque catégorie d'incident.

Une autre zone permet d'indiquer quelles garanties sont souscrites. Enfin, le code permet de visualiser les différents coefficients utilisés pour ajuster la fréquence : annualisation, effet du seuil de déclaration des données, adaptation à la France. Le tarificateur est donc facilement modifiable, éventuellement chaque année si de nouvelles études permettent d'affiner l'évolution temporelle des incidents et le nombre d'incidents en France.

La prime pure est ensuite estimée par Monte Carlo : 1 000 000 de simulations sont effectuées pour la fréquence de chaque incident, pour la probabilité de déclenchement de chaque garantie, toutes deux par une loi binomiale car systématiquement inférieures à 1, et enfin le même nombre de simulations est réalisé pour le nombre d'individus affectés à chaque incident. Pour chaque simulation, la survenue ou non d'un incident, le déclenchement ou non d'une garantie et le coût de la garantie sont multipliés, afin d'obtenir un coût si un incident est survenu et qu'il a déclenché une garantie.

Pour chaque coût simulé associé à un incident et à une garantie, la limite indiquée au contrat est appliquée. Les coûts de chaque garantie puis de chaque incident sont ensuite sommés afin d'appliquer

la franchise globale, ainsi que la limite globale annuelle.

La moyenne du vecteur de 1 000 000 valeurs obtenues après toutes ces étapes permet d'aboutir à la prime pure.

4.7 Comparaison des résultats avec le tarificateur actuel pour différents profils d'établissements

La section suivante présente les résultats obtenus comparés avec les primes pures issues du tarificateur actuel. Par souci de confidentialité, les primes présentées sont le résultat des primes réelles multipliées par un coefficient de passage arbitraire.

Pour différents profils d'établissements, la prime du modèle actuel est calculée pour un risque élevé, c'est à dire avec un questionnaire de risque informatique jugé passable, et pour un bon risque avec un questionnaire qui révèle une très bonne sécurité informatique. Cela indique ainsi pour un établissement donné une fourchette entre une prime basse et une prime haute selon leur état de santé informatique. Les primes pures sont calculées pour un contrat avec les garanties de base, ainsi qu'un contrat avec toutes les garanties.

Les profils sont classés par ordre croissant de chiffre d'affaires.

Garanties sélectionnées	Modèle actuel		Modèle mémoire
	Prime basse	**Prime haute**	**Prime modélisée**
Profil 1			
Garanties de base	2 686	6 260	2 691
Toutes garanties	3 691	8 133	2 843
Profil 2			
Garanties de base	6 879	16 475	1 124
Toutes garanties	9 492	21 420	1 577
Profil 3			
Garanties de base	8 357	19 563	7 412
Toutes garanties	11 288	25 367	8 422
Profil 4			
Garanties de base	10 671	25 193	19 687
Toutes garanties	13 587	31 409	22 074
Profil 5			
Garanties de base	43 392	110 198	39 760
Toutes garanties	63 687	151 756	51 934
Profil 6			
Garanties de base	77 631	203 529	144 784
Toutes garanties	118 198	287 926	185 375
Profil 7			
Garanties de base	73 870	195 976	152 673
Toutes garanties	113 727	279 160	194 453
Profil 8			
Garanties de base	126 459	323 992	716 680
Toutes garanties	165 065	414 733	939 715
Profil 9			
Garanties de base	229 175	580 726	1 123 119
Toutes garanties	283 365	716 347	1 513 321
Profil 10			
Garanties de base	199 592	502 508	533 871
Toutes garanties	300 274	722 716	739 882

TABLE 4.15 – Primes modèle actuel et modélisées dans ce mémoire pour différents profils d'établissement

De manière globale, le comportement du modèle construit dans ce mémoire est similaire à celui du modèle actuel : les primes augmentent avec la taille de l'établissement, mesurée par les variables Chiffre d'affaires et Nombre d'employés. Une analyse plus détaillée pour chaque profil est effectuée ci-dessous.

D'abord, les primes obtenues sont contenues dans la fourchette de primes proposée par le modèle actuel pour les profils 4, 6, 7, ainsi que pour le profil 1 avec les garanties de base uniquement.

Concernant le profil 3, le profil 5, et le profil 1 avec toutes les garanties, les primes sont légèrement plus faibles que celles du modèle actuel. A l'inverse, les primes du profil 10 sont légèrement plus élevées. Le chiffre d'affaires des profils 1, 3 et 5 est inférieur à 260 millions d'euros, contrairement au profil 10. Or, avoir un chiffre d'affaires supérieur à ce seuil va fortement augmenter la probabilité de divulgations, qui est l'incident le plus fréquent dans la modélisation. La discrétisation des variables aboutit en effet à

des résultats plus extrêmes que le modèle actuel qui semble avoir une hausse plus progressive des primes avec le chiffre d'affaires. De manière générale, les primes obtenues pour les petits établissements sont un peu plus faibles que celles du modèle actuel, et un peu plus élevées pour les grands établissements.

Enfin, des résultats très différents sont obtenus pour les profils 2, 8 et 9. Une analyse permet de mettre en valeur que ces écarts sont liés à la prise en compte différente de certaines variables :

— La prime obtenue pour le profil 2 est nettement en dessous de celle estimée par le modèle actuel. Une comparaison avec le profil 1 est effectuée pour illustrer la source de cette différence importante. Par rapport au profil 1, le profil 2 a un chiffre d'affaires plus élevé mais beaucoup moins d'ordinateurs et de serveurs. Le modèle actuel semble donc considérer que le profil 2 est un risque plus important en raison de son chiffre d'affaires. Inversement, le modèle construit au sein du groupe Relyens considère que le profil 2 est plus risqué que le profil 1. Même si la hausse du chiffre d'affaires entraîne une hausse modérée de la probabilité et du coût, elle ne compense pas la forte baisse par rapport au profil 1 du coût espéré associé au faible nombre d'ordinateurs et de serveurs, variable non prise en compte dans le modèle actuel.

— Les primes des profils 8 et 9 sont très supérieures à celles du modèle actuel. Pour la plupart des profils, la marge brute n'est pas renseignée dans les questionnaires et une valeur par défaut en pourcentage du chiffre d'affaires est indiquée. Les deux profils ont renseigné une marge brute plus faible que la valeur par défaut.
Le tableau suivant résume les primes après remplacement de la marge brute par la valeur par défaut dans le modèle du mémoire :

	Prime marge brute réelle	**Prime marge brute par défaut**	**Variation**
Profil 8			
Garanties de base	716 680	390 530	-46%
Toutes garanties	939 715	502 636	-47%
Profil 9			
Garanties de base	1 123 119	511 894	-54%
Toutes garanties	1 1513 321	672 663	-55%

TABLE 4.16 – Variation des primes profils 8 et 9 avec la marge brute par défaut

Remplacer la marge brute par la valeur par défaut entraîne une forte baisse de la prime allant de 46 à 55%. Au contraire, les primes du modèle actuel augmentent en moyenne de 10% en indiquant la marge brute par défaut pour ces profils. Les primes obtenues en appliquant la marge brute par défaut sont ainsi beaucoup plus cohérentes par rapport à celles du modèle actuel.

Dans la modélisation du mémoire, la marge brute est une variable qui baisse la probabilité de piratage et de piratage de prestataires lorsqu'elle augmente, ce qui entraîne une forte baisse de la prime. Une interprétation est que la marge brute est un indice de bonne santé financière et bonne gestion de l'établissement. Plus elle est élevée, plus l'établissement dispose de moyens financiers pour investir dans la sécurité informatique et dans la formation des employés. Cette variable n'est pas utilisée dans le coût du modèle construit. A l'inverse, le modèle actuel considère que la hausse de la marge brute est un risque plus élevé, car dans le contexte d'une entreprise de manière générale, en cas de déclenchement d'une garantie associée à la perte d'exploitation, le montant à indemniser serait plus élevé.

Ainsi, malgré des similitudes, des spécificités du secteur de la santé additionnelles sont observables.

Il faut également noter que les primes du modèle actuel présentées peuvent être ajustées par les souscripteurs du groupe Relyens dans une certaine proportion afin de s'adapter au mieux au risque.

4.8 Considérations opérationnelles et limites du modèle

Le modèle construit permet d'obtenir une prime pure pour le contrat proposé par Relyens, calculée à partir des informations financières et structurelles de l'établissement de santé, selon les garanties sélectionnées. Cependant, il ne prend pas en compte le niveau de sécurité informatique mesurable par des questionnaires. Or, dans le modèle actuel, le niveau de sécurité informatique influence beaucoup la prime. Par exemple, elle est au minimum multipliée par deux en passant d'un très bon risque à un mauvais risque. Le modèle pourra à l'avenir être complété par d'autres variables explicatives mesurant cette dimension si des données venaient à être disponibles, ou bien par un ajustement de la prime prenant en compte le niveau de sécurité.

Ensuite, la modélisation de la probabilité se base sur des données peu nombreuses, les résultats ont dû être validés avec les experts en risque informatique afin de s'assurer qu'il n'y avait pas d'effets fortuits. De plus, la discrétisation des variables quantitatives mène à des résultats très segmentés. Les données sinistres pourront être complétées au fur et à mesure, et les données contrats mises à jour afin d'apporter plus d'informations au modèle.

Il faut également noter que de nombreux ajustements ont été nécessaires pour adapter la modélisation de la probabilité à la France. Ces ajustements pourront facilement être mis à jour annuellement. Cependant, ils n'ont permis qu'une adaptation imparfaite au contexte français ; en particulier, il a été supposé que la proportion de chaque type d'incident était la même entre les Etats-Unis et la France, et que dans les deux pays la divulgation était l'incident le plus fréquent, ce qui n'est pas forcément le cas.

N'ayant pas les coûts associés aux incidents modélisés, ils ont été déterminés avec l'aide des experts en risque informatique. Ces coûts sont prudents, des études régulières pourront être réalisées afin de les ajuster, et une modélisation pourra être mise en place si des données viennent à être disponibles.

Enfin, le risque Cyber se caractérise par un risque d'accumulation important du fait de l'interconnexion entre les établissements. Les variables Membre d'un système et Logiciel externe de données n'ont pas été considérées comme augmentant la probabilité d'incident dans la modélisation. Cependant, le risque est bien présent. Le fournisseur de logiciel *Eye Care Leaders* a subi un piratage en 2022 impactant les données patients de plusieurs ophtalmologues. [46]. Le risque d'accumulation est donc bien réel et peut faire l'objet d'études supplémentaires à l'avenir.

46. Source : https://www.govinfosecurity.com/cloud-based-ehr-vendor-hack-affects-eye-care-practices-a-19066

En résumé, les modèles de probabilité de survenance de chaque incident sont rééstimés de la même manière que la partie précédente, à partir uniquement des variables disponibles et comparables avec un établissement français. Il est possible d'observer que pour les trois incidents modélisés, la variable chiffre d'affaires est la plus influente et augmente la probabilité de survenue. La probabilité est ensuite annualisée, puis assimilée à une fréquence car très peu d'établissements ont déclaré plusieurs incidents. La fréquence est ajustée en fonction du seuil de déclaration de données, et du nombre d'incidents informatiques subis par les établissements de santé en France.

Afin de finaliser le modèle de la forme fréquence $\times$ coûts, une matrice de coûts est construite afin d'associer à chaque type d'incident un coût par garantie, ainsi qu'une probabilité de déclenchement des garanties. Un outil est mis en forme afin d'obtenir une prime à partir des informations renseignées, par simulation de Monte Carlo.

Enfin, les primes sont calculées pour différents profils réels d'établissements, et leur valeur sont comparées avec celles du modèle actuel. Il est possible d'observer un comportement similaire des deux modèles, hormis pour deux profils d'établissement. Les différences s'expliquent principalement par l'incorporation de variables plus liées au périmètre à risque comme le nombre de serveurs et d'ordinateurs. Elles dépendent également de la prise en compte différente de certaines variables notamment la marge brute.

Ainsi, le modèle construit permet d'obtenir une prime pure prenant en compte le risque de survenance et le coût de plusieurs types d'incidents Cyber. En raison du faible nombre de données, les résultats sont à prendre avec prudence et le modèle pourra être amélioré, par exemple par un enrichissement périodique de la base des incidents et une mise à jour des différents facteurs d'ajustements.

Conclusion

La tendance actuelle de hausse des incidents Cyber, et en particulier des cyberattaques, s'observe à la fois aux Etats-Unis et en France, et n'épargne pas le secteur de la santé. Les médias informent de plus en plus fréquemment de la survenue d'une nouvelle cyberattaque perturbant le fonctionnement d'un établissement de soins, et le nombre de déclarations en France par les établissements de santé a doublé entre 2020 et 2021. Il a été montré dans ce mémoire qu'aux Etats-Unis, les hôpitaux ont connu une hausse linéaire dans le temps des piratages portant atteinte aux données personnelles de santé depuis 2014, ainsi qu'une très forte hausse des piratages de prestataires à partir de 2020.

Les modèles de régression logistique estiment la probabilité pour un hôpital de déclarer un incident en se basant sur la période 2014 à 2021. Pour chaque incident, le meilleur modèle est recherché en comparant différentes méthodes de rééchantillonnage. Les modèles retenus parviennent à mettre en avant certains profils d'hôpitaux américains plus touchés par les incidents étudiés. Plusieurs variables sont en effet influentes selon le test de rapport de vraisemblance, et leurs modalités sont influentes selon le test de Wald. Les valeurs de l'aire sous la courbe ROC, ainsi que les résultats des tests d'Hosmer–Lemeshow valident les modèles, mais il faut noter que les prédictions sont faites sur la même base que celle ayant servi à la modélisation. Les valeurs de l'aire sous la courbe ROC estimées par validation croisée montrent un ajustement satisfaisant, mais les bonnes prédictions des établissements restent limitées comme le montre les valeurs obtenues de F1-Score. D'autres effets sont probablement impliqués dans la probabilité de survenance comme la sécurité informatique de l'établissement, mais ces données sensibles ne sont pas accessibles dans les bases de données publiques.

Un travail de sélection des variables de la base contrats est réalisé afin de conserver uniquement celles comparables à des données disponibles pour un établissement en France. Certaines variables comme les Liquidités, les Impayés et les Revenus des soins ne sont pas disponibles en France et ne peuvent donc pas être utilisées par Relyens pour estimer le risque de ses sociétaires. De plus, les mesures de comptabilité françaises et américaines ne sont pas toujours comparables. Les modèles de régression logistique sont à nouveau estimés en incluant toutes les variables qui sont utilisables par l'entreprise, puis en sélectionnant les variables influentes.

Les modèles permettent d'estimer une probabilité d'incident Cyber. L'étude de l'évolution temporelle des différents incidents permet d'écrire des équations entre les différentes probabilités annuelles dans le but d'annualiser la probabilité estimée par les modèles sur toute la période. Cette probabilité est ensuite assimilée à une fréquence, ce qui permet de lui appliquer différents ajustements afin de se rapprocher du contexte en France.

Un indicateur de l'ampleur des incidents présent dans les données est le nombre d'individus affectés. Pour chaque type d'incident, une loi est ajustée par la méthode des moments, et est validée par le test de Kolmogorov-Smirnov. Le nombre d'individus touchés par un incident est ensuite utilisé pour calculer une partie du coût associé à un incident, complété par d'autres informations, qui permettent la construction d'une matrice de coûts.

La matrice indique des coûts associés à chaque garantie, pour chaque catégorie d'incident et éventuellement à une maille plus fine comme le type d'évènement (rançongiciel, hameçonnage, ..). Elle permet de finaliser le modèle fréquence × coûts. La fréquence de chaque incident est simulée par Monte Carlo, ainsi que la probabilité de déclenchement de chaque garantie et le nombre d'individus affectés. Le coût est ainsi calculé selon le nombre de personnes touchées et les caractéristiques de l'établissement.

Les primes sont ensuite calculées avec le modèle construit pour différents profils, et comparées avec les résultats fournis par le modèle actuel. Les différentes primes obtenues révèlent une tendance similaire entre les deux modèles : une hausse du chiffre d'affaires et du nombre d'employés accroît le risque. Les primes sont néanmoins moins lissées que celles du modèle actuel, avec des résultats proches de la tranche basse des primes pouvant être proposées par le modèle actuel pour les petits établissements, et des primes proches de la tranche haute pour les gros établissements. De plus, certains profils ont permis de mettre en évidence des divergences sur la prise en compte de certaines informations : le modèle construit prend en compte d'autres variables pour mesurer la taille d'un établissement de santé qui n'est pas toujours directement lié au niveau du chiffre d'affaires. Également, une interprétation contraire de la variable marge brute est observée : une forte marge augmente le risque d'après le modèle actuel, alors qu'elle baisse la probabilité de survenance d'un incident d'après la modélisation.

Ainsi, le modèle construit permet d'obtenir des résultats de primes qui sont à prendre avec précaution. En effet, en raison du faible nombre de données, la robustesse des modèles de probabilité d'incident Cyber est limitée. De plus, la discrétisation des variables quantitatives mène à des résultats segmentés. Concernant les coûts des incidents, ils ne sont pas connus et ont donc dû être estimés par avis d'expert. Une autre limite est l'absence de variables mesurant directement le niveau de sécurité informatique, alors qu'il s'agit d'une dimension qui influence beaucoup le risque. Enfin, les hypothèses choisies et les ajustements appliqués ne permettent d'aboutir qu'à une adaptation imparfaite au contexte français. Ce modèle est donc une première proposition de modélisation qui vise à être développée, par exemple avec l'ajout des données sinistres qui seront publiées dans les années à venir, et l'intégration d'une analyse supplémentaire portant sur le niveau de sécurité informatique.

L'interconnexion entre les établissements est ici prise en compte par les variables Membre d'un système et Logiciel externe de données dans les modèles de probabilité. Ces variables se sont révélées non ou peu significatives. Il s'agit cependant d'une source de risque qui pourrait faire l'objet d'études supplémentaires, en étudiant le risque d'accumulation propre au secteur de la santé, dont les établissements sont particulièrement interconnectés en étant membres de structures communes, en mutualisant leurs serveurs, ou bien en utilisant les mêmes logiciels.

Table des figures

Liste des tableaux

Bibliographie

AHA (American Hospital Association). *AHA Hospital Statistics*, 2020 edition.

AMRAE. *LUCY (LUmière sur la CYberassurance)*, Etudes 2021 et 2022.

BEAUD DE BRIVE G. Modélisation du risque cyber pour un portefeuille d'assurance français. *Institut des Actuaires*, 2021.

CARAVAGNA L. Sécurité informatique : les établissements de santé désignés ose ont une double obligation de signalement des incidents (anssi), 2021. https://www.ticsante.com/story?ID=5747.

CRO Forum. *Cyber resilience. The cyber risk challenge and the role of insurance*, 2014.

DEBES F. Cybersécurité : le plan à 1 milliard de l'etat, 2021. Article publié sur https ://www.lesechos.fr/tech-medias/hightech/cybersecurite-le-plan-a-1-milliard-de-letat-1291369.

DELCAMP C. *Les risques numériques et la cyberassurance.* Support de la formation "La gestion des risques cyber de la prevention à la couverture assurantielle" délivrée par l'Institut du Risk Management en Juin 2021.

DIXNEUF P. Analyse de la performance de la méthode d'imputation de données manquantes miss-forest et application à des données environnementales. *Mémoire de maîtrise électronique, Montréal, École de technologie supérieure*, 2019.

DJOFFON O. Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel. *Institut des Actuaires*, 2017.

EIOPA. *Cyber Risk for Insurers - Challenges and Opportunities*, 2019.

EIOPA. *Understanding Cyber Insurance*, 2017.

ELING M. et SCHNELL W. Ten key questions on cyber risk and cyber risk insurance. *The Geneva Association*, 2016.

Groupe de Travail IA. *Emergence du besoin en cyber assurance. Institut des Actuaires*, 2017.

IBM. *Cost of a Data Breach Report*, 2021.

KOWARIK A. et TEMPL M. Imputation with the r package vim. *Journal of Statistical Software*, 74(7) :1–16, 2016.

LARSEN Kim. *Data Exploration with Weight of Evidence and Information Value in R.* https://multithreaded.stitchfix.com/blog/2015/08/13/weight-of-evidence/.

Liptak Z. *The q-gram distance.* Support de cours disponible sur http://profs.scienze.univr.it/~liptak/FundBA/slides_1718/StringDistance2_6up.pdf.

Liptak Z. *The q-gram distance*. Support de cours disponible sur http://profs.scienze.univr.it/~liptak/FundBA/slides/StringDistance2_6up.pdf.

MANSKI, C. F., LERMAN, S. R. The estimation of choice probabilities from choice based samples. *Econometrica, 45*, 1977.

MARTINEZ A. Modélisation assurantielle du risque cyber. *Institut des Actuaires*, 2019.

Ministère des solidarités et de la santé. *Observatoire des signalements d'incidents de sécurité des systèmes d'information pour le secteur santé*, Rapport 2021.

Ministère des solidarités et de la santé, Agence du numérique en santé. *Cybersécurité dans le secteur de la santé et du médico-social*, Dossier d'information 2021.

NAIC. *Cybersecurity Insurance Market*, 2021.

PLANCHET F. et MISERAY A. *Tarification IARD Introduction aux techniques avancées*, 2017.

PONS F. Etude actuarielle du cyber risque. *Institut des Actuaires*, 2014.

RAKOTOMALALAl R. *Pratique de la Régression Logistique Régression Logistique Binaire et Polytomique*, 2015.

ROUVIERE Laurent. Régression logistique avec r. Support de cours disponible sur https://lrouviere.github.io/doc_cours/poly_logistique.pdf.

ROUVIERE Laurent. Analyse du modèle de régression logistique. Support de cours disponible sur https://perso.univ-rennes2.fr/system/files/users/rouviere_l/chapitre2_glm.pdf.

STEKHOVEN Daniel J. et BÜHLMANN Peter. *MissForest - nonparametric missing value imputation for mixed-type data*, 2011.

ZIRAR W. *Le coût total de la cyberattaque du CH de Dax s'est élevé à 2,3 millions d'euros (RSSI)*, 2022. https://www.ticsante.com/story?ID=6141.

XIE,Y. MANSKI, F. The logit model and response-based samples. *Sociol. Methods Res., 17(3) :283–302*, 1989.

ZHANG Lili, RAY Herman, PRIESTLEY Jennifer , TAN Soon. A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, 2018.

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. (*Incidents de violations de données de santé de plus de 500 individus*).

https://data.cms.gov/. (*Données publiées par CMS*).

https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports/Cost-Reports-by-Fiscal-Year. (*Données des Cost Reports*).

https://www.ssi.gouv.fr/administration/. (*Site de l'ANSSI décrivant les menaces Cyber et les réglementations*).

https://www.hhs.gov/hipaa/for-professionals/breach-notification/breach-reporting/index.html. *(Règles de notification des violations de données personnelles de santé).*

www.ahrq.gov/chsp/data-resources/compendium.html. (*Source de données et d'informations sur les systèmes aux Etats-Unis*).

https://www.healthit.gov/data/datasets/ehr-products-used-meaningful-use-attestation. (*Source de données sur les fournisseurs de logiciels de données de santé*).

https://www.hipaajournal.com/what-are-the-penalties-for-hipaa-violations-7096/14. (*Amendes en cas de non respect de la réglementation HIPAA*).

https://privacyrights.org/resources/health-insurance-portability-and-accountability-act. (*Evolution de la réglementation HIPAA aux Etats-Unis*).

https://www.fda.gov/medical-devices/postmarket-requirements-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities. (*Réglementation sur les dispositifs médicaux aux Etats-Unis*).

https://data.drees.solidarites-sante.gouv.fr/explore/dataset/708_bases-statistiques-sae/information/. (*Base de données statistiques de la DRESS*).

https://donnees-rgpd.fr/loi-informatique-libertes/. (*Loi informatique et libertés*).

https://solidarites-sante.gouv.fr/professionnels/gerer-un-etablissement-de-sante-medico-social/groupements-hospitaliers-de-territoire/. (*Informations sur les GHT*).

https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/ce-que-l-on-sait-de-la-cyberattaque-contre-l-hopital-de-villefranche-sur-saone_4299065.html. (*Article sur l'attaque ayant touché l'hôpital de Villefranche-sur-Saône*).

https://www.francetvinfo.fr/internet/securite-sur-internet/cyberattaques/cyberattaque-a-l-hopital-de-corbeil-essonnes-patients-et-membres-du-personnel-incites-a-porter-plainte_5407198.html. (*Article sur l'attaque ayant touché l'hôpital de Corbeil-Essonnes*).

https://france3-regions.francetvinfo.fr/occitanie/herault/montpellier/chu-montpellier-victime-attaque-informatique-plus-600-ordinateurs-infectes-1670717.html. (*Article décrivant la cyberattaque subie par l'hôpital de Montpellier*).

https://www.ihs.gov/bemidji/healthcarefacilities/casslake/. (*Site internet de présentation d'un hôpital américain*).

https://www.ticsante.com/story?ID=4715. (*Article décrivant la condamnation par la CNIL de l'hôpital Haga de la Haye*).

https://www.cybermalveillance.gouv.fr/. (*Site internet gouvernemental de sensibilisation et d'assistance aux incidents Cyber*).

https://www.cnil.fr/fr/thematique/cnil/sanctions. (*Site de la CNIL répertoriant les sanctions prononcées*).

https://www.cnil.fr/fr/cnil-direct/question/quels-sont-les-grands-principes-des-regles-de-protection-des-donnees. (*Site de la CNIL présentant la RGPD*).

https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance. (*Distance de Damerau Levenshtein*).

https://fr.wikipedia.org/wiki/DistancedeJaro-Winkler. (*Distance de Jaro-Winkler*).

https://www.r-bloggers.com/2021/04/robservations-9-the-hosmer-lemeshow-test-follows-a-chi-square-distribution-water-is-wet/. (*Présentation du test d'Hosmer-Lemeshow*).

https://www.lesechos.fr/tech-medias/hightech/cybersecurite-le-plan-a-1-milliard-de-letat-1291369. (*Article décrivant le plan de cybersécurité national en 2021*).

https://www.govinfosecurity.com/cloud-based-ehr-vendor-hack-affects-eye-care-practices-a-19066. (*Article décrivant le piratage subi par le prestataire Eye Care Leaders en 2022*).

https://fr.wikipedia.org/wiki/Distance_de_Jaro-Winkler. (*site décrivant la distance de Jaro-Winkler*).

https://en.wikipedia.org/wiki/DameraužLevenshtein_distance. (*site décrivant la distance de Damerau-Levenshtein*).

https://www.r-bloggers.com/2021/04/robservations-9-the-hosmer-lemeshow-test-follows-a-chi-square-distribution-water-is-wet/. (*site décrivant le test d'Hosmer-Lemeshow*).

Annexes

Annexe A

Glossaire Acronymes

Abbréviation	Signification
ACM	Analyse par correspondances multiples
ACP	Analyse en composantes principales
AHA	*American Hospital Association*
AMRAE	Association pour le Management des Risques et des Assurances de l'Entreprise
ANS	Agence du Numérique en Santé
ANSSI	Agence Nationale de la Sécurité des Systèmes d'Information
ARS	Agence Régionale de Santé
CE	Conformité européenne
CERT	*Computer Emergency Response Team*
CHU	Centre Hospitalier Universitaire
CMS	*Centers for Medicare and Medicaid Services*
CNIL	Commission Nationale de l'Informatique et des Libertés
CRO	Chief Risk Officers
DPI	Dossier Patient Informatisé
EHR	*Electronic Health Record*
EIOPA	*European Insurance and Occupational Pensions Authority*
ENISA	*European Union Agency for Cybersecurity*
FDA	*U.S. Food and Drug Administration*
FSN	Fournisseurs de Services Numériques
GHT	Groupements Hospitaliers de Territoire
HHS	*U.S. Department of Health and Human Services*
HIPAA	*Health Insurance Portability and Accountability Act*
HITECH	*he Health Information Technology for Economic and Clinical Health*
ISO	*International Standards Organization*
IT	*Information Technology*
LUCY	LUmière sur la CYberassurance
NAIC	*National Association of Insurance Commission*
NIS	*Network and Information System Security*
NRMSE	Racine de l'erreur quadratique moyenne
OIV	Opérateur d'Importance Vitale
ONC	Office of the National Coordinator for Health Information Technology
OSE	Opérateur de Services Essentiels
OT	*Operational Technology*
PFC	Proportion de données mal classées
RGPD	Règlement Général pour La Protection des Données
SaaS	*Software-as-a-Service*
WOE	*Weight of Evidence*

Annexe B

Fouille de textes (*Text Mining*)

B.1 Nuage de mots de la variable Description

.

Les 100 mots les plus fréquents apparaissant dans la variable description sont présentés sur ce nuage de mots [47].

FIGURE B.1 – Nuage de mots de la variable Description

47. après retrait des mots *and, the, that, its, this, all, with, center, which, experienced, also, other, from, were, two, had, was, for, inc, dba*

B.2 Mots clés recherchés pour la création de variables supplémentaires

Dans cette section est détaillée pour chaque variable les mots-clés recherchés pour associer une modalité.

— Type de donnée

Modalité	Mots clés associés
Démographique	*name,addresses,dates of birth, Social Security numbers, driver's license, ssn, other identifier, demographic*
Clinique	*health insurance information, treatment information, lab results, diagnosis, conditions, medications, clinical, medical record*
Financier	*claims information, credit card, bank account, financial information*

TABLE B.1 – Mots clés utilisés pour créer la variable type de données

— Sous-catégorie de l'incident

Modalité	Mots clés associés
Rançongiciel	*ransomware,ransom*
Hameçonnage	*phishing, mail scam, malicious email*
Malware	*malware, malicious, spy, virus*
Erreur non intentionnelle	*error,inadvertently, unsecure manner, unsecured email, failed, emailed an invitation, erroneously, unintended, incorrect e-mail, blind carbon copy, blind copy, employee lost, unintentional, mistake, accidentally, inadvertently, wrong recipients, improperly,security incident, misconfigured, not password protected, configuration, server was taken offline, vulnerability*
Accès non autorisé	*viewed medical records,impermissibly accessed, without a necessary business reason, impermissibly electronically accessed, accessed the protected, had been accessing, impermissibly acquired, accessed patient, nurse accessed, accessed medical, innappropriately accessed, accessed protected, employee had been accessing, employees accessed, impermissibly obtained, employee accessed, employee removed, impermissibly access,credentials, unauthorized third party, unauthorized individual*
Négligence	*impermissibly emailed, impermissibly sent, personal mail account,personal email account, personal website, personal and professional email account, home email account,without notifying patients, without authorization, impermissibly transmitted, impermissibly disclosed,impermissibly used*

TABLE B.2 – Mots clés utilisés pour créer la variable Sous-catégorie de l'incident

De plus pour cette variable les modalités Vol/Perte d'un ordinateur de bureau/objet informatique/autre sont construites à partir des variables type d'incident et localisation.

— Réponse à l'incident

TABLE B.3 – Mots clés utilisés pour créer la variable réponse à l'incident

Modalité	Mots clés associés
Formation des employés	*retrained, re-trained, training*
Assistance technique	*technical assistance*
Sanction des employés	*sanction, employee was terminated, against the former employee*
Mise en place de nouvelles procédures de sécurité informatique	*risk management plan, risk analysis, revised policies and procedures, policy, updated policies and procedures, reinforce, procedure, policies*
Mise en place de nouvelles technologies de protection	*encryption technolog, encryption technolog, encryption,changed password, strengthened password, password, safeguard, additional technical safeguard, implemented safeguard, additional technical and security safeguard, additionnal safeguard, additional administrative technical and security safeguard, improved several of its safeguard, additionnal firewall safeguard procedure*
Prestations de protection contre l'usurpation d'identité et de surveillance du crédit	*theft restoration services, complimentary phone number, protection services, credit*
Investigations	*investigate,audit*
Changement de la relation avec un prestataire	*updated ba agreement, revised business associate contracts*
Analyse du risque	*risk analysis*
A bénéficié de l'assistance d'OCR	*technical assistance*
Changement des mots de passe	*changed password,strengthened password,password*
Mise en place des moyens de protection physique	*physical*
Adoption de technologies de chiffrage	*encryption technolog,encryption technolog,encryption*
Réalisation d'enquêtes	*investigate,audit*
Mise en place de nouveaux plans de gestion des risques	*risk management plan*

Fin du contrat avec le partenaire	*ceased sending, the ce terminated,ceased sending, the ce terminated, the ce had ceased doing business*
Implémentation d'évaluations périodiques	*periodic evaluation,periodic technical and nontechnical evaluation,evaluation*

Annexe C

Méthodes d'estimation de la similarité entre deux noms

C.1 Distance de Damerau-Levenshtein

Cette distance compte le nombre minimum d'insertions, de substitutions, de suppressions ou de transpositions de deux caractères adjacents pour transformer un mot en un autre.

La fonction de distance récursive entre le i^e symbole de a et le j^e symbole de b est définie par [48]

$$d_{ab}(i,j) = min \begin{cases} 0 & \text{si } i = j = 0, \\ d_{ab}(i-1,j) + 1 & \text{si } i > 0, \\ d_{ab}(i,j-1) + 1 & \text{si } j > 0, \\ d_{ab}(i-1,j-1) + \mathbb{1}_{a_i \neq b_j} & \text{si } i,j > 0, \\ d_{ab}(i-2,j-2) + 1 & \text{si } i,j > 1 \text{ et } a_i = b_{j-1} \text{ et } a_{i-1} = b_j \end{cases}$$

et la distance de Damerau-Levenshtein est $d_{ab}(|a|,|b|)$ avec $|.|$ la longueur d'un mot.

C.2 Distance de Jaro-Winkler

La distance de Jaro-Winkler mesure les correspondances entre caractères dont les positions dans les mots sont proches, et ajoute une pénalité lorsque les caractères sont transposés. [49]

Elle provient de la distance de Jaro qui vaut

$$d_j = \frac{1}{3}\left(\frac{m}{|a|} + \frac{m}{|b|} + \frac{m-t}{m}\right)$$

avec m le nombre de caractères correspondants et t le nombre de transpositions. Deux caractères sont correspondants si la différence entre leur position respective dans le mot est inférieure à :

48. Source : https://en.wikipedia.org/wiki/DamerauâĂŞLevenshtein_distance
49. Source : https://fr.wikipedia.org/wiki/Distance_de_Jaro-Winkler

$$\lfloor \frac{max\big(|a|,|b|\big)}{2} \rfloor - 1$$

Le nombre de transpositions désigne le nombre de fois que des caractères à la même position et correspondants sont différents, divisé par deux.

La distance de Jaro-Winkler a pour but de donner plus de poids à la ressemblance si le début des mots est le même. Elle vaut

$$d_w = d_j + (lp\,(1 - d_j))$$

avec l la longueur du préfixe < 4 et p un poids.

C.3 Distance N-gramme

Un n-gramme est un mot de longueur n. La distance n-gramme entre deux mots a et b se définit par [50] :

$$d\,(a,b) = \sum_{u \in Q} |N\,(a,u) - N\,(b,u)\,|$$

où Q représente l'ensemble des mots de longueur n formés à partir des lettres présentes dans a et b. $N(.,u)$ désigne le nombre de fois que le n-gramme u apparaît dans le mot.

C.4 Distance basée sur la sous séquence commune la plus longue

Les mots a et b sont comparés afin de déterminer la sous séquence commune aux deux mots la plus longue sans modification d'aucun caractère. La distance utilisée correspond au nombre de caractères qui ne se trouvent pas dans la sous séquence commune.

La fonction de distance compte le nombre de suppressions et insertions nécessaires pour transformer un mot en un autre.

50. Source : http://profs.scienze.univr.it/~liptak/FundBA/slides/StringDistance2_6up.pdf

Annexe D

Tests statistiques

D.1 Test d'indépendance du χ^2

Soit V et W deux variables qualitatives avec respectivement n et p modalités. N est le nombre d'individus. n_{ij} désigne le nombre d'individus avec la i^e modalité de V et la j^e modalité de W, $n_{i.}$ correspond au nombre d'individus avec la modalité i de V, $n_{.j}$ le nombre d'individus avec la modalité j de W. L'écart à l'indépendance entre V et W vaut

$$d^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{N})^2}{\frac{n_{i.}n_{.j}}{N}}$$

Il est à comparer avec la valeur du χ^2 pour un nombre de degrés de liberté de $(n-1)(p-1)$.

D.2 Test d'Hosmer-Lemeshow

Ce test a pour but d'estimer la qualité d'ajustement d'un modèle de régression binaire. [51]. Pour g groupes avec généralement $g = 10$, la statistique de test est

$$G = \sum_{j=1}^{g} \frac{(o_j - e_j)^2}{e_j} \sim \chi^2_{g-2}$$

où o_j est le nombre de succès observés, e_j est le nombre de succès espérés.

D.3 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est effectué sur deux échantillons afin de déterminer s'ils suivent la même loi.
Soit $(x_1, \ldots, x_n)$ un échantillon de taille n d'une variable aléatoire X de fonction de répartition F. La

51. https://www.r-bloggers.com/2021/04/robservations-9-the-hosmer-lemeshow-test-follows-a-chi-square-distribution-water-is-wet/ *(site décrivant le test d'Hosmer-Lemeshow)*

fonction de répartition empirique F_n de X est définie par :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq x} = \begin{cases} 0 & \text{if } x < x_1 \\ \frac{i-1}{n} & \text{if } x_{i-1} \leq x < x_i \\ 1 & \text{if } x \geq x_n \end{cases} \tag{D.1}$$

Si le premier échantillon est de taille n et de fonction de répartition F, le second échantillon de taille m et de fonction de répartition G.

Les deux hypothèses testées sont $H_0 : F = G$ versus $H_1 : F \neq G$.

La statistique du test est :

$$D_{n,m} = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|$$

Si X et Y suivent la même loi, $(\frac{1}{n} + \frac{1}{m})^{1/2} D_{n,m}$ suit une loi de Kolmogorov.

D.4 Test de Wald sur plusieurs coefficients

Le test de Wald a pour but de tester la nullité simultanée de q coefficients. Les hypothèses sont $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$ contre $H1 : \beta_1 = \beta_2 = \ldots = \beta_q \neq 0$.

La statistique du test suit une loi du χ^2 à q degrés de liberté, elle vaut :

$$W = \hat{\beta}'_{(q)} \times \hat{\Sigma}^{-1}_{(q)} \times \hat{\beta}_{(q)}$$

avec $\hat{\beta}_{(q)}$ le sous-vecteur des valeurs observées des coefficients testés, $\hat{\Sigma}^{-1}_{(q)}$ la sous-matrice de variance covariance associée à ces coefficients.

Annexe E

Liste des variables de la base contrats

Les variables de la base contrats sont sélectionnées depuis la base d'usage publique des déclarations financières des hôpitaux à *Medicare* et *Medicaid*. La colonne NA indique la proportion de valeurs manquantes pour chaque variable.

TABLE E.1 – Variables de la base contrats

Nom en français	Nom en anglais	Description	NA
Accueille des internes	*Teaching Hospital*	Hôpital qui a le droit à un remboursement car il participe au programme *GME (Graduate Medical Education)*	0,15
Actif courant	*Total Current Assets*	Valeur des actifs courants	0,04
Actif fixe	*Total fixed Assets*	Valeur de tous les actifs immobilisés	0,06
Actif Total	*Total Assets*	Valeur des actifs courants, des actifs immobilisés, et des autres actifs	0,05
Adresse	*Street Address*	Adresse postale	0
Autres recettes	*Total Other Income*	Autres revenus que ceux des hospitalisations et consultations	0,05

Canton	*County*		0
Certification données	*Meaningful use*		0,15
Charges patients Ambulatoire	*Outpatient Total Charges*	Prix des soins aux patients en ambulatoire	0.08
Charges patients hospitalisés	*Inpatient Total Charges*	Prix des soins aux patients hospitalisés	0.01
Charges patients	*Combined Outpatient Inpatient Charges*	Somme des deux variables précédentes	0.01
Code postal	*Zip Code*		0
Coût dépréciation	*Depreciation Costs*	Dépréciation, amortissement	0.02
Coûts hors salaires	*Overhead Non Salary Costs*	Dépenses hors salaires	0.01
Coût total	*Total Costs*	Coûts totaux	0.01
Date de début de la période fiscale	*Fiscal Year Begin Date*		0
Date de fin de la période fiscale	*Fiscal Year End Date*		0
Dépenses opérationnelles	*Less Total Operating Expense*	Dépenses opérationnelles ou d'exploitation	0,01

Etat	*State Code*		0
Hôpital d'accès critique	*Critical Access Hospital*	Hôpital de petite taille dans une zone rurale, ayant pour but de faciliter l'accès aux soins	0,15
Impayés	*Total Bad Debt Expense*	Somme des montants dûs par des patients qui ne sont pas payés	0,24
Liquidités	*Cash on Hand and in Banks*	Montant immédiatement disponible pour un financement	0,08
Nom de l'hôpital	*Hospital Name*		0
Nombre d'internes	*Number of Interns and Residents FTE*	Pour calculer l'ETP (Equivalent Temps Plein), le temps de travail est ajusté à celui d'un temps plein pour les internes	0,77
Nombre de lits activité principale	*Number of Beds*	Comprend les lits physiquements disponibles, occupés ou non, de l'unité principale de l'établissement. Variable utilisée pour l'imputation des valeurs manquantes mais pas pour la modélisation.	0,01
Nombre de lits	*Number of Beds Total for all subproviders*	Comprend les lits physiquements disponibles, occupés ou non, de toutes les unités de soins de l'hôpital.	0,01
Nombre employés ETP	*FTE Employees on Payroll*	Pour calculer l'ETP (Equivalent Temps Plein), le temps de travail est ajusté à celui d'un temps plein	0,02
Nombre de journées activité principale	*Total Days V XVIII XIX Unknown*	Somme du nombre de jours de soins pour les patients de l'activité principale de l'établissement pendant la période. Variable utilisée pour l'imputation des valeurs manquantes mais pas la modélisation.	0,01

Nombre de journées	*Total Days V XVIII XIX Unknown Total for all subproviders*	Somme du nombre de jours de soins pour tous les patients pendant la période	0,01
Nombre de jours-lits activité principale	*Total Bed Days V XVIII XIX Unknown*	Nombre de lits multiplié par le nombre de jours de la période, en prenant en compte les variations du nombre de lits présents pendant la période. Variable utilisée pour l'imputation des valeurs manquantes mais pas pour la modélisation.	0,01
Nombre de jours-lits	*Total Bed Days V XVIII XIX Unknown Total for all subproviders*	Nombre de lits multiplié par le nombre de jours de la période, en prenant en compte les variations du nombre de lits présents pendant la période pour toutes les unités de soins	0,01
Nombre de sorties activité principale	*Total Discharges V XVIII XIX Unknown*	Patients sortis de l'hôpital ou décédés des unités de soins principales. Variable utilisée pour l'imputation des valeurs manquantes mais pas pour la modélisation.	0,01
Nombre de sorties	*Total Discharges V XVIII XIX Unknown Total for all subproviders*	Patients sortis de l'hôpital ou décédés de toutes les unités de soins	0,01
Numéro de déclaration	*rpt_rec_num*	Numéro assigné pour chaque déclaration	0
Numéro d'identification *CCN*	*Provider CCN*	*CMS Certification number (CCN)* identifiant chaque établissement	0

Montant Salaires	*Total Salaries From Worksheet A*	Montant des salaires versés	0.01
Passif courant	*Total Current Liabilities*	Valeur du passif courant	0,04
Passif long terme	*Total Long Term Liabilities*	Valeur du passif immobilisé	0,19
Passif Total	*Total Liabilities*	Somme des deux variables précédentes	0,05
Résultat net	*Net Income*	Concerne toutes les activités de l'établissement	0,01
Résultat Net des Soins	*Net Income from Service to Patients*	Revenu réel des soins - Dépenses opérationnelles	0,01
Résultat total	*Total Income*	Resultat net patients + Revenus autre que ceux des patients	0,01
Ratio Coûts Charges	*Cost To Charge Ratio*	Charges des patients / Coûts totaux de l'établissement	0,22
Revenus Soins Ambulatoires	*Outpatient Revenue*	Revenus issus de l'activité en ambulatoire	0,12
Revenus hospitalisations	*Inpatient Revenue*	Revenus issus de l'activité au sein de l'établissement	0,05
Revenu brut	*Gross Revenue*	Somme des deux variables précédentes	0,04
Revenu Soins	*Net Patient Revenue*	Recettes effectives, après ajustement d'éventuelles réductions	0,04

Rural ou Urbain	*Rural Versus Urban*	Indicateur géographique	0,01
Service d'urgences	*Emergency Services*		0,12
Solde Général	*General Fund Balance*		0,05
Type de contrôle	*Type of Control*	Plusieurs catégories d'organisations à but non lucratif, gouvernementales ou privées	0
Type de soins	*Provider Type*	Soins aigus, de long terme, psychiatriques, de rééducation, pour enfants, contre le cancer, contre les addictions ou dans une institution religieuse non médicale	0,14
Type d'établissement selon la classification de *CMS* (TypCMS)	CCN Facility Type	Accès critique, enfants, long terme, psychiatrique, réhabilitation, court terme, institution religieuse non médicale	0
Valeur bâtiments	*Buildings*	Tous les coûts associés aux bâtiments	0,17
Ville	*City*		0
Valeur équipement mobile	*Major Movable Equipment*	Coût de l'équipement mobile d'une valeur de plus de 5 000$ utilisé dans les opérations, de durée de vie généralement supérieure à 3 ans.	0,14